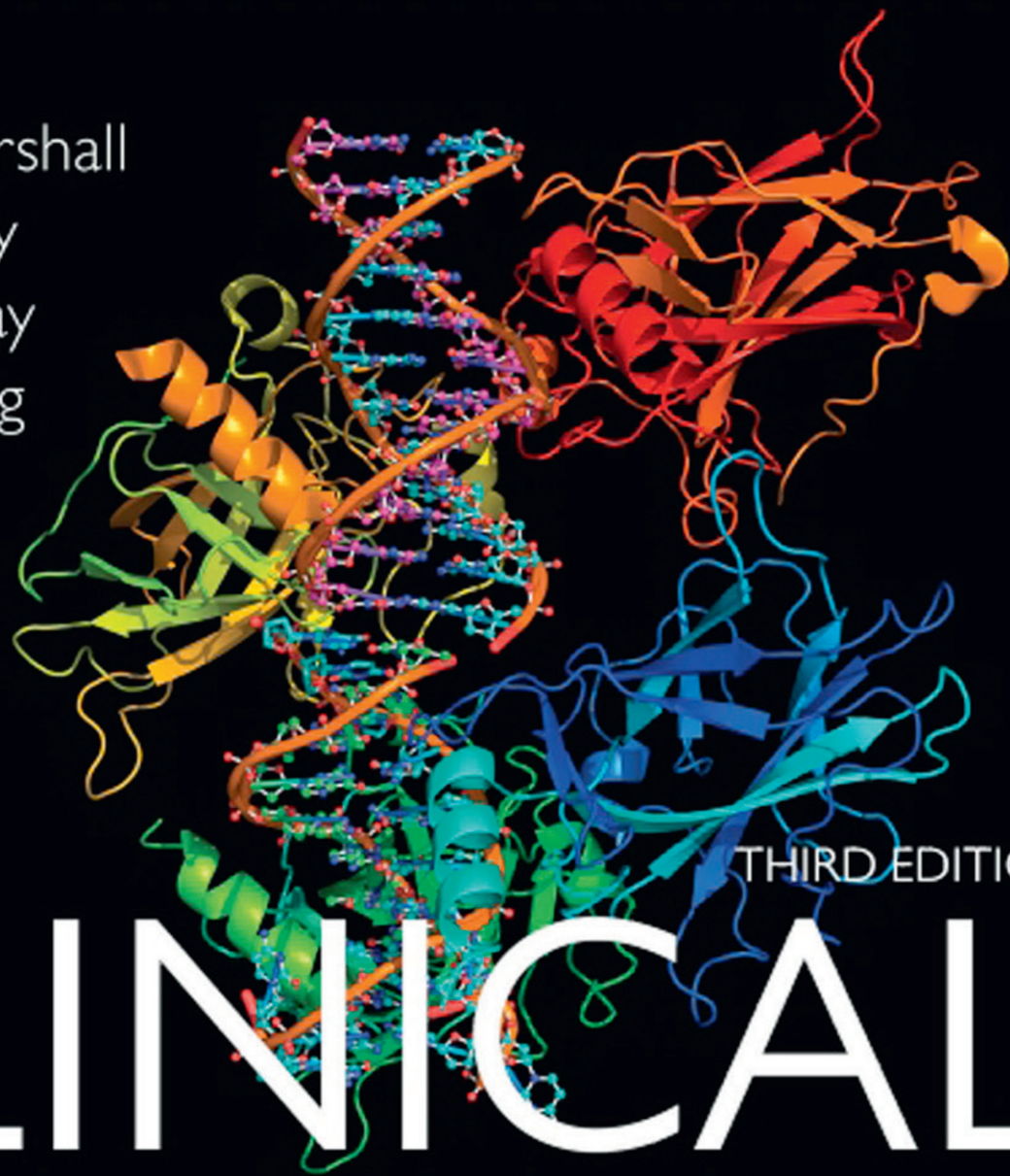


William J. Marshall
Marta Lapsley
Andrew P. Day
Ruth M. Ayling



THIRD EDITION

CLINICAL

BIOCHEMISTRY

METABOLIC AND CLINICAL ASPECTS

CHURCHILL
LIVINGSTONE
ELSEVIER



CLINICAL
BIOCHEMISTRY
Metabolic and clinical aspects

For Elsevier:
Content Strategist: *Jeremy Bowes*
Senior Content Development Specialist: *Ailsa Laing*
Senior Project Manager: *Beula Christopher*
Designer: *Miles Hitchen*
Illustrator: *Chartwell Illustrations*

CLINICAL BIOCHEMISTRY

Metabolic and clinical aspects

THIRD EDITION

EDITED BY

**William J. Marshall MA PhD MSc MB BS FRCP
FRCPATH FRCPedin FRSC FSB FLS**

Consultant Clinical Biochemist and Clinical Director of Pathology,
The London Clinic;
Emeritus Reader in Clinical Biochemistry,
King's College London,
London, UK

Marta Lapsley MB BCh BAO MD FRCPATH

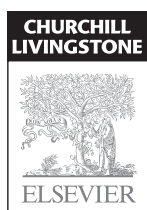
Consultant Chemical Pathologist,
Epsom and St Helier University Hospitals NHS Trust,
London, UK;
Honorary Senior Lecturer in Clinical Endocrinology and Nutrition,
University of Surrey,
Surrey, UK

Andrew P. Day MA MSc MB BS FRCPATH

Consultant Chemical Pathologist,
Weston Area Health Trust and University Hospitals Bristol Foundation Trust;
Honorary Senior Clinical Lecturer in Chemical Pathology,
University of Bristol,
Bristol, UK

Ruth M. Ayling BSc MB BS MSc PhD FRCP FRCPATH

Consultant Chemical Pathologist,
Derriford Hospital,
Plymouth, UK



Edinburgh London New York Oxford Philadelphia St Louis Sydney Toronto 2014

An imprint of Elsevier Limited

© 2014 Elsevier Limited. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

First edition 1995
Second edition 2008
Third edition 2014

ISBN 978-0-7020-5140-1
eBook ISBN 978-0-7020-5478-5

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data

A catalog record for this book is available from the Library of Congress

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

With respect to any drug or pharmaceutical products identified, readers are advised to check the most current information provided (i) on procedures featured or (ii) by the manufacturer of each product to be administered, to verify the recommended dose or formula, the method and duration of administration, and contraindications. It is the responsibility of practitioners, relying on their own experience and knowledge of their patients, to make diagnoses, to determine dosages and the best treatment for each individual patient, and to take all appropriate safety precautions.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ELSEVIER your source for books,
journals and multimedia
in the health sciences
www.elsevierhealth.com



The
Publisher's
policy is to use
paper manufactured
from sustainable forests

Contents

- Preface vii
- Contributors viii
- 1 Uses of biochemical data in clinical medicine 1
William J. Marshall • Marta Lapsley
 - 2 Acquisition and interpretation of biochemical data 6
Helen Bruce • Marta Lapsley
 - 3 Quality aspects of laboratory medicine 21
Helen Bruce • Marta Lapsley
 - 4 Sodium, water and potassium 27
Michael D. Penney
 - 5 Hydrogen ion homeostasis and tissue oxygenation and their disorders 65
William J. Marshall
 - 6 Calcium, phosphate and magnesium 93
Timothy Cundy • Andrew Grey • Ian R. Reid
 - 7 The kidneys, renal function and kidney disease 124
David Makanjuola • Marta Lapsley
 - 8 Proteinuria 152
Anne Dawnay
 - 9 Renal tubular disorders and renal stone disease 168
David Makanjuola • Marta Lapsley
 - 10 Clinical biochemistry of nutrition 180
Ruth M. Ayling
 - 11 Nutritional disorders and their management 200
Ruth M. Ayling
 - 12 Clinical biochemistry of the gastrointestinal tract 214
Ingvar T. Bjarnason • Roy A. Sherwood
 - 13 Assessment of hepatic function and investigation of jaundice 231
Roy A. Sherwood • Adrian Bomford
 - 14 Acute and chronic liver disease 250
Adrian Bomford • Roy A. Sherwood
 - 15 Glucose metabolism and the pathophysiology of diabetes mellitus 273
David B. Wile • John P.H. Wilding
 - 16 The clinical management of diabetes mellitus 305
Ian W. Seetho • John P.H. Wilding
 - 17 Hypoglycaemia 333
Mourad H. Labib
 - 18 Hypothalamic, pituitary and adrenal disorders 349
Miles J. Levy • Trevor A. Howlett
 - 19 Thyroid dysfunction 373
Colin M. Dayan • Onyebuchi E. Okosieme • Peter Taylor
 - 20 Metabolic response to stress 403
Robin Berry • Philip Gillen
 - 21 Disorders of puberty and sex development 412
S. Faisal Ahmed • Jane D. McNeilly
 - 22 Reproductive function in the female 433
Leslie D. Ross
 - 23 Reproductive function in the male 451
John Miell • Zoe Davies
 - 24 Inherited metabolic disease 461
Fiona Carragher • Mike Champion
 - 25 Paediatric clinical biochemistry 484
Fiona Carragher
 - 26 Introduction to haematology and transfusion science 497
David Ah-Moye • Ceinwen Davies • Joanne Goody • Peter Hayward • Rebecca Frewin
 - 27 Biochemical aspects of anaemia 515
Rebecca Frewin
 - 28 The porphyrias: inherited disorders of haem synthesis 533
Michael N. Badminton • George H. Elder
 - 29 The haemoglobinopathies 550
David C. Rees • Roopen Arya
 - 30 Immunology for clinical biochemists 560
Joanna Sheldon • Rachel D. Wheeler • Pamela G. Riches

- 31** Metabolic bone disease 604
Timothy Cundy • Ian R. Reid • Andrew Grey
- 32** Biochemistry of articular disorders 636
Jeremy G. Jones
- 33** Muscle disease 646
Laurence A. Bindoff
- 34** Investigation of cerebrospinal fluid 660
Geoffrey Keir • Carrie Chadwick
- 35** Biochemical aspects of psychiatric disorders 673
William J. Marshall • Teifion Davies
- 36** Biochemical aspects of neurological disease 683
Paul Hart • Clare M. Galtrey • Dominic C. Paviour • Min Htut
- 37** Lipids and disorders of lipoprotein metabolism 702
Graham R. Bayly
- 38** Clinical biochemistry of the cardiovascular system 737
Clodagh M. Loughrey • Ian S. Young
- 39** Therapeutic drug monitoring 767
Mike Hallworth
- 40** Poisoning 787
James W. Dear
- 41** Metabolic effects of tumours 808
Wassif S. Wassif • James E. East
- 42** Tumour markers 821
Catharine M. Sturgeon
- 43** Molecular clinical biochemistry 844
Roberta Goodall
- 44** Forensic biochemistry 874
Robert J. Flanagan • Sarah Belsey • Terhi Launiainen
- Index 883

Preface

In the preface to the second edition of this book, two major changes in the practice of clinical biochemistry were noted. The first was the increasing integration between the pathology disciplines, driven largely by shared technology and now reflected in the multi-disciplinary training for many healthcare scientists. The second was the increasing tendency for medically qualified clinical biochemists to have direct responsibility for the management of patients with metabolic diseases. Both these trends have continued, and we have recognized them in preparing this third edition. The sections on haematology and immunology have been expanded so that, while not attempting to provide detailed accounts of these subjects, we believe that we have provided sufficient information to allow clinical biochemists to familiarize themselves with both their laboratory and clinical aspects, and to be in a position to seek greater knowledge from specialist textbooks as required.

The aspects of metabolic disease for which medical clinical biochemists may have responsibility include nutritional disorders, diabetes, inherited metabolic disease (particularly in adults), metabolic bone disease, renal calculi and dyslipidaemias, and we have encouraged our authors to provide sufficient detail to convey the general principles of the diagnosis and management of these conditions. We believe that this material will also be of interest to scientist clinical biochemists, by helping to set the more scientific material in its clinical context.

The overall aim of the book remains unchanged: to provide, in a single volume, a textbook of clinical biochemistry both for senior trainees and for established practitioners. We have not included details of analytical methodology, which are well covered in other books, but have included a new chapter devoted to quality management, since this is such an important topic: laboratory data are useless – and potentially dangerous – if their quality cannot be assured. Although the processes of quality management are to a considerable extent centred on the laboratory, they start and finish with patients and their medical attendants.

Comments from reviewers of the previous edition have encouraged us to include material that is usually outside the scope of textbooks of clinical biochemistry. New chapters on the metabolic response to stress and forensic aspects of clinical biochemistry increase the breadth of

coverage in a way that we hope readers will find useful, informative and relevant.

We thank our numerous contributors, old and new, for their commitment to this project, their (in most cases) adherence to deadlines and their tolerance of our editorial input. No one writes book chapters for the money; we are grateful to them all for their ready acceptance of our invitations and for their time.

The greatest change with this edition has been the recruitment of three new editors – Ruth Ayling, Andrew Day and Marta Lapsley (lead editor) who have joined William Marshall in place of Stephen Bangert, whose other commitments precluded his being involved. Two editors took the main responsibility for each chapter, but all of us read and commented on all the material, and approved the final versions. We hope that we will thereby have produced an error-free manuscript and apologize for any errors that may have slipped through. Will any reader who spots one please let us know, so that it can be corrected at reprinting?

At Elsevier, Jeremy Bowes commissioned the project, but day-to-day management has once again been in the capable hands of Ailsa Laing. We are indebted to her for her constant encouragement and for liaising with our authors with regard to the delivery of their manuscripts. Not to have had to do this in addition to sorting out editorial matters has been a huge help. Our thanks also go to the in-house team, particularly Beula Christopher, who coordinated the typesetting and proof corrections, and the designers and the copy editor, who have transformed the multitude of different styles of documents and images into a coherent finished product that is a pleasure to handle and read.

Most importantly of all, we would also like to mention the long suffering families, friends and colleagues who have listened to endless discussions about the work involved in editing the book and have provided practical support over many months to relieve us of more mundane tasks. In particular we would like to thank Wendy (Marshall), Michèle (Day) and Michael (Lapsley) who have contributed significantly, albeit indirectly, to the final publication.

Ruth M. Ayling
Andrew P. Day
Marta Lapsley
William J. Marshall

Contributors

S. Faisal Ahmed, MB ChB MD FRCPCH

Professor of Child Health, Honorary
Consultant in Paediatric Endocrinology,
University of Glasgow, Royal Hospital
for Sick Children, Glasgow, UK

David Ah-Moye, HND(MLS) FIBMS DMLM

Biomedical Scientist (Deputy Laboratory Manager),
Haematology Department, Gloucestershire Royal
Hospital, Great Western Road, Gloucester, UK

Roopen Arya, BM BCh MA PhD FRCP FRCPATH

Consultant Haematologist, Department of
Haematological Medicine, King's College Hospital,
London, UK

Ruth M. Ayling, BSc MB BS MSc PhD FRCP FRCPATH

Consultant Chemical Pathologist, Derriford Hospital,
Plymouth, UK

Michael N. Badminton, MBChB PhD FRCPATH

Senior Lecturer and Honorary Consultant,
Department of Medical Biochemistry and
Immunology, School of Medicine, Cardiff
University, Cardiff, UK

Graham R. Bayly, BA BM FRCP FRCPATH

Consultant Biochemist, Bristol Royal Infirmary,
Bristol, UK

Sarah Belsey, BSc MSc

Clinical Scientist, Toxicology Unit, Department of
Clinical Biochemistry, King's College Hospital,
London, UK

Robin Berry, PhD FRCA DICM

Consultant in Anaesthetics and Intensive Care,
Derriford Hospital, Plymouth, UK

Laurence A. Bindoff, MSc MRCP MD

Professor of Neurology, Department of Clinical
Medicine, University of Bergen, Haukeland
University Hospital, Bergen, Norway

Ingvar T. Bjarnason, MD MSc FRCPATH**FRCP(Glasg) DSc**

Professor of Digestive Diseases, Department of
Gastroenterology, King's College Hospital,
London, UK

Adrian Bomford, MD FRCP

Reader in Medicine/Honorary Consultant Physician,
Institute of Liver Studies, King's College Hospital,
London, UK

Helen Bruce, MChem MSc FRCPATH

Principal Clinical Scientist, Department of Clinical
Biochemistry, Royal Surrey County Hospital,
Surrey, UK

Fiona Carragher, MSc FRCPATH

Consultant Clinical Scientist, Department of Chemical
Pathology, Guy's and St Thomas' NHS Foundation
Trust, London, UK

Carrie Chadwick, BSc (Hons) MSc FRCPATH

Consultant Clinical Scientist, Aintree University
Hospital Foundation Trust; Laboratory Director,
The Walton Centre Foundation Trust,
Liverpool, UK

Mike Champion MSc FRCP FRCPCH

Consultant in Paediatric Metabolic Medicine,
Department of Paediatric Metabolic Medicine,
Evelina Children's Hospital, Guy's and St Thomas'
NHS Foundation Trust, London, UK

Timothy Cundy, MA MD FRCP FRACP

Professor of Medicine, Department of Medicine,
Faculty of Medical and Health Sciences, University of
Auckland, New Zealand

Ceinwen Davies, BSc MSc

Coagulation Supervisor, Haematology Department,
Gloucestershire Royal Hospital, Gloucester, UK

**Teifion Davies, BSc MB BS PhD DPMSA MSB CBiol
FRCPsych**

Director of Undergraduate Psychiatry Teaching,
King's College London, Institute of Psychiatry,
London, UK

Zoe Davies, MA BM BCh MRCP

Specialist Trainee in Diabetes and Endocrinology,
London Deanery, London, UK

Anne Dawnay, PhD FRCPATH

Consultant Biochemist and Honorary Senior Lecturer,
University College Hospitals, London, UK

Colin M. Dayan, MA MB BS FRCP PhD

Professor of Clinical Diabetes and Metabolism,
Director, Institute of Molecular and Experimental
Medicine, Cardiff University School of Medicine,
Cardiff, Wales, UK

James W. Dear, PhD FRCPedin

Consultant in Clinical Pharmacology, National
Poisons Information Service, Royal Infirmary of
Edinburgh; Senior Clinical Lecturer, Queen's
Medical Research Institute, University of
Edinburgh, Edinburgh, UK

James E. East, BSc MB ChB MRCP MD

Consultant Gastroenterologist,
John Radcliffe Hospital, Oxford, UK

George H. Elder, MD FRCP FRCPATH FMedSci

Emeritus Professor, Department of Medical
Biochemistry and Immunology, School of Medicine,
Cardiff University, Cardiff, UK

**Robert J. Flanagan, PhD ERT MFSSoc CChem FRSC
FRCPATH HFCMHP**

Consultant Clinical Scientist, Toxicology Unit, Clinical
Biochemistry, King's College Hospital, London, UK

Rebecca Frewin, BSc MB ChB MRCP FRCPATH

Consultant Haematologist, Haematology Department,
Edward Jenner Unit, Gloucester Royal Hospital,
Gloucester, UK

Clare M. Galtrey, MA MB BChir PhD MRCP

Specialist Trainee in Neurology, St George's Hospital,
London, UK

Philip Gillen, FRCA FFICM

Consultant in Anaesthetics and Intensive Care,
Derriford Hospital, Plymouth, UK

Roberta Goodall, BSc MSc FIMLS FRCPATH

Formerly Consultant Scientist, Department of Clinical
Biochemistry, North Bristol NHS Trust, Southmead
Hospital, Bristol, UK

Joanne Goody, BSc MSc

Assistant Laboratory Manager, Blood Transfusion
Department, Gloucestershire Hospital NHS
Foundation Trust, Cheltenham General Hospital,
Cheltenham, UK

Andrew Grey, MD FRACP

Associate Professor of Medicine, Department of
Medicine, Faculty of Medical and Health Sciences,
University of Auckland, New Zealand

Mike Hallworth, MA MSc FRCPATH EurClinChem

Consultant Clinical Scientist, Royal Shrewsbury
Hospital, Shrewsbury, UK

Paul Hart, MB BS BSc FRCP PhD

Consultant Neurologist, Epsom and St. Helier NHS
Trust, London; Atkinson Morley Neuroscience
Unit, St Georges Hospital; Royal Marsden Hospital,
London, UK

Peter Hayward, BSc MSc

Blood Transfusion Section Supervisor,
Gloucestershire Hospital NHS
Foundation Trust, Cheltenham General Hospital,
Cheltenham, UK

Trevor A. Howlett, MD FRCP

Consultant Physician and Endocrinologist, Department
of Diabetes and Endocrinology, Leicester Royal
Infirmary, Leicester, UK

Min Htut, MB BS MMedSci MD MRCP DGM DTM&H

Consultant Neurologist and Honorary Consultant
Neurophysiologist, Epsom and St Helier University
Hospitals NHS Trust, London, and St. George's
Hospital, London, UK

Jeremy G. Jones, MD FRACP FAFRM

Consultant Rheumatologist, North West Wales NHS
Trust, Ysbyty Gwynedd; Senior Clinical Lecturer,
School of Sport, Health and Exercise Sciences,
University of Wales, Bangor, UK

Geoffrey Keir, PhD MSc FRCPATH FIBMS

Clinical Scientist and Honorary Senior Lecturer,
Neuroimmunology and CSF Laboratory, Institute of
Neurology, The National Hospital for Neurology and
Neurosurgery, London, UK

Mourad H. Labib, MB ChB FRCPATH

Consultant Chemical Pathologist, Clinical Biochemistry
Department, Dudley Group of Hospitals, Russells
Hall Hospital, West Midlands, UK

Marta Lapsley, MB BCh BAO MD FRCPATH

Consultant Chemical Pathologist, Epsom and
St Helier University Hospitals NHS Trust, London, UK;
Honorary Senior Lecturer in Clinical Endocrinology and
Nutrition, University of Surrey, Surrey, UK

Terhi Launiainen, PhD

Postdoctoral Researcher, Department of Forensic
Medicine, Hjelt Institute, Faculty of Medicine,
University of Helsinki, Helsinki, Finland

Miles J. Levy, MD FRCP

Consultant Physician and Endocrinologist, Department
of Diabetes and Endocrinology, Leicester Royal
Infirmary, Leicester, UK

Clodagh M. Loughrey, MD MRCP FRCPATH

Consultant Chemical Pathologist, Belfast Trust, Belfast,
Northern Ireland

Jane D. McNeilly, BSc(Hons) MSc PhD FRCPPath
Clinical Biochemist, Biochemistry Department,
Southern General Hospital Laboratory,
Glasgow, UK

David Makanjuola, MD, FRCP
Consultant Nephrologist, Department of Renal
Medicine, Epsom and St. Helier University Hospitals
NHS Trust, London, UK

**William J. Marshall, MA PhD MSc MB BS FRCP FRCPPath
FRCPedin FRSC FSB FLS**
Consultant Clinical Biochemist and Clinical Director
of Pathology, The London Clinic; Emeritus Reader
in Clinical Biochemistry, King's College London,
London, UK

John Miell, DM FRCP FRCPEdin
Divisional Director of Specialty Medicine and
Pathology and Head of Metabolic Medicine
Department, Consultant Physician and
Endocrinologist, University Hospital Lewisham,
London, UK

Onyebuchi E. Okosieme, MD FRCP
Diabetes and Endocrinology Department, Prince
Charles Hospital, Cwm Taf Health Board,
Merthyr Tydfil, UK

Dominic C. Paviour, BSc PhD MB BS MRCP
Consultant Neurologist and Honorary Senior
Lecturer, Epsom and St Helier Hospitals NHS
Trust and St George's University of London,
London, UK

Michael D. Penney, BSc MB BS MD FRCPPath
Consultant Chemical Pathologist, Department of
Clinical Biochemistry, Royal Gwent Hospital,
Gwent, UK

David C. Rees, MA MB BS FRCP FRCPPath
Senior Lecturer in Paediatric Haematology,
Department of Paediatric Haematology,
King's College London School of Medicine,
King's College Hospital, London, UK

Ian R. Reid, BSc MB ChB MD FRACP FRCP FRSNZ
Professor of Medicine and Endocrinology,
Department of Medicine, Faculty of Medical and
Health Sciences, University of Auckland, New
Zealand

Pamela G. Riches, HNC BSc PhD FRCPPath
Retired Professor, Department of Immunology,
St George's Hospital, London, UK

**Leslie D. Ross, MB BS MRCS LRCP FRCS(Ed)
FRCOG DMI**
Consultant Obstetrician and Gynaecologist, Epsom and
St Helier University Hospitals NHS Trust, London, UK

Ian W. Seetho, BA BMedSci BM BS MRCP
Clinical Research Fellow, Department of Obesity and
Endocrinology, Clinical Sciences Centre, University
Hospital Aintree, Liverpool, UK

Joanna Sheldon, PhD CSci MRCPPath FRCPPath
Consultant Immunologist and Honorary Senior
Lecturer, Department of Immunology, St George's
Hospital, London, UK

Roy A. Sherwood, BSc MSc DPhil
Professor of Clinical Biochemistry, King's College
London; Consultant Biochemist, King's College
Hospital, London, UK

Catharine M. Sturgeon, BSc PhD FRCPPath
Clinical Scientist, Department of Clinical Biochemistry,
Royal Infirmary of Edinburgh; Honorary Fellow,
University of Edinburgh, Edinburgh, UK

Peter Taylor, BSc MB ChB MRCP MSc
Welsh Clinical Academic Trainee in Diabetes and
Endocrinology, Thyroid Research Group, Institute
of Molecular and Experimental Medicine, School of
Medicine, Cardiff University, Cardiff, UK

**Wassif S. Wassif, MB ChB MSc CSci Eur Clin Chem MD
FRCP FRCPPath**
Consultant Clinical Biochemist and Head of
Department, Clinical Biochemistry, Bedford Hospital
NHS Trust, Bedford, UK

Rachel D. Wheeler, MA MSc PhD
Clinical Scientist in Immunology, Protein Reference
Unit, St George's Hospital, London, UK

John P.H. Wilding, DM FRCP
Professor of Medicine and Honorary Consultant
Physician; Head of Department of Obesity and
Endocrinology, Institute of Ageing and Chronic
Disease, Clinical Sciences Centre, University Hospital
Aintree, Liverpool, UK

David B. Wile, MB BCh MSc FRCPPath
Associate Specialist in Clinical Biochemistry, Diabetes
and Endocrinology, Department of Clinical
Biochemistry, University Hospital Aintree NHS
Foundation Trust, Liverpool, UK

Ian S. Young, BSc MB BCh MD FRCP FRCPPath
Professor of Medicine and Director of the Centre for
Public Health, Queen's University Belfast, Belfast,
Northern Ireland

Uses of biochemical data in clinical medicine

William J. Marshall • Marta Lapsley

CHAPTER OUTLINE

INTRODUCTION 1

SPECIFIC USES OF BIOCHEMICAL TESTS 2

Diagnosis 2

Management 3

Screening 4

Other uses of biochemical investigations 5

CONCLUSION 5

INTRODUCTION

The science of biochemistry is fundamental to the practice of clinical medicine. Many diseases have long been known to have a biochemical basis and research in biochemistry is increasingly providing descriptions of pathological processes and explanations for disease at a molecular level.

As a result of the application of biochemical principles and techniques to the analysis of body fluids and tissues, clinicians have an extensive and ever-increasing range of biochemical investigations that can be called upon to aid clinical decision-making. Such investigations can provide information vital to the diagnosis and management of many conditions, including both those with an obvious metabolic basis (e.g. diabetes mellitus) and those in which metabolic disturbances occur as a consequence of the disease (e.g. renal failure). On the other hand, many conditions are successfully diagnosed and treated without recourse to any biochemical investigation, while there remain conditions in which it might be expected that biochemical investigations should be of value but for which appropriate tests are not yet available. For example, there are, as yet, no practical biochemical investigations to assist in the diagnosis and management of the major affective disorders (see Chapter 35), although there is considerable evidence that biochemical disturbances are involved in the pathogenesis of these conditions.

Biochemical analysers range from large, automated instruments capable of performing multiple tests on single serum samples to relatively simple instruments designed to measure only one or a few analytes. In general, they generate results quickly, reliably and economically. However, some tests, often more complex and expensive ones, are performed manually and may take longer to complete. Biochemical data are thus readily available to support clinical decision-making. Ordering a biochemical investigation is a simple procedure and there is no doubt that such investigations are

often requested automatically, without regard for their potential value in the specific clinical setting. Clinical biochemists decry this but do themselves no favour by their use of the term, widely employed by clinicians, 'routine investigations' (usually meaning relatively simple investigations that are performed frequently) and even 'routine laboratories' (meaning the places where they are done).

Ideally, investigations should always be performed because there is a specific indication for them, that is, because it is anticipated that their results will provide information of benefit to the management of the patient. However, it cannot be denied that investigations requested for no specific reason can sometimes provide valuable information. Most clinicians are able to recall occasions when an unexpected result from a 'routine test' has provided the essential clue to the diagnosis in a difficult case. More often, the finding of an unexpectedly abnormal result may engender considerable anxiety and involve further investigations to elucidate its cause, only for it to transpire that the biochemical abnormality is of no clinical significance.

The potential range of investigations available to support the clinician is considerable, from simple low cost urine dip-stick tests to magnetic resonance imaging using hugely expensive equipment. There is an understandable tendency for clinical biochemists to think that biochemical investigations are pre-eminent among special investigations. In some conditions they are, in others they have no role, while in many, their value is greatly increased when their results are considered alongside those of other investigations, for example imaging. The clinician should be aware of the whole range of investigations that are available, but needs also to be able to appreciate their various advantages and limitations. The clinical biochemist, too, needs to be aware of the role of other investigations, so that he or she can view biochemical tests in context and advise on their suitability and the interpretation of their results in specific clinical circumstances.

It has been the editors' aim to ensure that this information is provided where relevant in this book.

The processes of acquiring and interpreting biochemical data are complex. Correct interpretation requires that the clinical context and reason for requesting the test are properly understood, otherwise the result has little value. This chapter explores the variety of potential ways in which biochemical data can be used in clinical practice.

SPECIFIC USES OF BIOCHEMICAL TESTS

Diagnosis

It has been said that diagnosis in medicine is an art, not a science, yet the process of diagnosis is amenable to scientific analysis. Making a diagnosis is the equivalent of propounding a hypothesis. A hypothesis should be tested by experiment, the results of which may support or refute the hypothesis, which can then be extended, modified or discarded in favour of an alternative, as appropriate. The validity of a clinical diagnosis is tested by observation of the natural history of the condition or its response to appropriate treatment or by the results of definitive investigations: the diagnosis will be confirmed if these are as expected from knowledge of previous cases. If they are not, it must be reviewed.

Clinical diagnosis is based on the patient's history and clinical examination. Taking general and hospital practice together, it has been estimated that, in more than 80% of cases, a confident diagnosis can be made on the basis of the history or the history and clinical findings alone. Even when this cannot be done, it should be possible to formulate a differential diagnosis, that is, a list of diagnoses that could explain the clinical observations. The results of investigations may then lead to one of these being considered the most likely and providing a rational basis for treatment. Subsequent observation will indicate whether the diagnosis was correct.

Although not necessarily required for the management of an individual patient, it may be possible to extend the clinical diagnosis by further investigation to determine the pathogenesis of the condition and ultimately, its underlying cause. For example, measurement of serum troponin concentration may confirm a clinical diagnosis of myocardial infarction in a patient with typical chest pain and electrocardiographic abnormalities; angiography could be used to demonstrate coronary atherosclerosis prior to surgery or angioplasty; the finding of hypercholesterolaemia would indicate a causative factor for the atherosclerosis; a family history of premature heart disease would suggest that the hypercholesterolaemia was familial and DNA mutation analysis might reveal the underlying genetic defect.

The ideal diagnostic investigation would be 100% sensitive (all cases of the condition in question would be correctly diagnosed) and 100% specific (no individual without the condition would be wrongly diagnosed as having it). The concepts of specificity and sensitivity are examined fully in Chapter 2. In practice, the capacity of biochemical investigations to provide precise diagnostic information is extremely variable. At one end of the spectrum, the techniques of genetic analysis are making it possible to reliably diagnose inherited metabolic diseases

in utero; at the other end of the spectrum, to take just one example, a decrease in plasma sodium concentration can occur in many different conditions and is, on its own, diagnostic of none of them.

Molecular genetic analysis has become a separate discipline in its own right and is a special case for the use of biochemical investigations for diagnosis. It is used to detect the presence of a mutation responsible for a specific disease. Even when possession of a mutation does not inevitably result in the development of a disease, its presence can indicate increased susceptibility to a condition. However, even individuals with the same genotype for a characteristic may differ in their phenotypes. But, although molecular genetics is a rapidly developing field, many genetically determined conditions, including inherited metabolic diseases, are still diagnosed on the basis of their biochemical phenotype.

With the exception of genetically determined diseases, the number of conditions in which biochemical investigations alone provide a precise diagnosis is very small. There are several reasons for this. First, biochemical changes are often a consequence of a pathological process that is common to many conditions. Thus, although tissue destruction leads to the release of intracellular enzymes into the plasma, few such enzymes are specific to any one tissue and tissue destruction can occur for many reasons, for example with ischaemia, exposure to toxins etc. Second, it also frequently happens that a biochemical variable can be influenced by more than one type of process. To cite a familiar example, plasma albumin concentration can be influenced by changes in the rates of synthesis and degradation of the protein and by changes in its volume of distribution, and the rate of synthesis in turn depends on substrate supply and hepatic function, among other factors. Third, even when a biochemical change is specific to one condition, it may not indicate its cause and this may need to be established before the condition can be treated appropriately. For example, the demonstration of a high plasma concentration of the thyroid hormone, tri-iodothyronine, is characteristic of hyperthyroidism, but this can be a result of several different thyroid diseases, and treatment appropriate for one of these may not be appropriate for another.

When a biochemical investigation is used for diagnosis, the result obtained from the patient will usually be compared with a reference range, that is, the range of values that can be expected in comparable apparently healthy individuals. The theory of reference ranges is discussed further in Chapter 2, but two points require particular emphasis here.

First, the natural variation of biochemical parameters is such that the ranges of concentrations of constituents of the plasma are likely to be narrower in an individual than in a group (even if well-matched to the individual). Second, for many biochemical variables, there is overlap, often considerable, between the range of values seen in healthy individuals and those characteristic of disease. Thus, a test result in a patient with a disease may fall into the range typical for healthy people and vice versa. This overlap stems in part from the fact that some organs have considerable reserve capacity. The liver, kidneys, pancreas and small intestine exemplify this. For example, in chronic kidney disease, renal function may still be sufficient to

maintain normal homeostasis with respect to body fluid composition, even when half the functional capacity of the kidneys has been lost. It should not, therefore, be surprising that simple measurements of function can yield normal results in patients with kidney disease. In chronic pancreatitis, biochemical evidence of functional disturbance (e.g. of malabsorption) usually only becomes apparent when at least 80% of the functional capacity of the pancreas is lost, although the characteristic of severe pain often occurs at an earlier stage. Similarly, disease of the small intestine by no means always results in malabsorption.

When previous measurements are available in an individual, test results can be compared with these values, rather than with a reference range. Indeed, biochemical investigations are sometimes requested to provide a 'baseline' against which to assess future results, particularly if there is risk of a particular complication developing or if a change can be anticipated from the natural history of the disease or the expected response to treatment. A change in a biochemical variable in relation to a previous result may be of significance, even if both results are within the reference range. Thus, a rise of creatinine concentration within the reference range may indicate a significant loss of kidney function, and may even indicate acute kidney injury if the rise has occurred rapidly. The concept of the critical difference between two results is discussed further in Chapter 2.

The capacity of a biochemical test to provide diagnostic information can be quantified by the calculation of a mathematical function known as the predictive value. As will be discussed in Chapter 2, the predictive value of a diagnostic test depends on the prevalence of the condition in the group of people to whom the test is applied. If a diagnostic test is used indiscriminately, its predictive value will be low.

The majority of biochemical investigations made for clinical purposes involve analysis of plasma or serum. However, changes in the concentration of analytes in these fluids do not necessarily parallel changes in intracellular or whole body content, and yet it may be these quantities that are more relevant to the underlying pathology. Furthermore, single measurements may not provide reliable information in non-steady-state situations, for example, the plasma concentration of thyroid stimulating hormone (TSH), which is typically very low in patients with thyrotoxicosis, may remain low for some weeks after treatment has rendered patients clinically euthyroid.

For whatever purpose biochemical data are used, it is essential that they are reliable and are available in time to be of use. Under some circumstances, it may be permissible to sacrifice some quality in order to obtain a result rapidly, but in general, every attempt should be made to minimize the influence of both analytical and preanalytical factors on the accuracy and precision of data. This topic is considered further in Chapter 2.

Management

Assessment of disease severity

Most biochemical investigations are quantitative, and the more abnormal a result is, the more likely it is that a pathological disturbance is causing it. Often, the extent

to which a result is abnormal correlates well with the severity of a condition but this is not always the case. The diagnostic test may not reflect that aspect of the condition of greatest importance in terms of severity; thus two patients with hepatitis may have equally raised plasma aminotransferase activities (reflecting tissue damage), but the condition will be judged more severe if, in one patient, the prothrombin time is prolonged (reflecting impaired hepatic functional capacity). Furthermore, overall disease severity (in relation to its effect on the patient) is likely to depend on many other factors, including the nature of the condition itself, the patient's age, previous state of health, the existence of other illness etc. For example, hepatitis due to infection with hepatitis A virus tends to have a good prognosis compared with that caused by hepatitis B or C.

Prognosis

In general, the results of biochemical tests are poor indicators of prognosis, but there are exceptions to this. For example, the plasma bilirubin concentration at the time of diagnosis in patients with primary biliary cirrhosis correlates well with outcome; a high plasma concentration of α -fetoprotein in a patient with testicular teratoma is of prognostic significance, but the concentration of paraprotein in a patient with myeloma is not. Other examples are discussed in the ensuing chapters.

One aspect of prognosis is the assessment of the benefits and risks of treatment. It has long been appreciated that different patients do not necessarily respond identically to the same drug. Many factors impinge on patients' responses to drugs including, for example, nutrition or the concomitant use of other drugs. Genetic factors are also important, and often many different genes are involved. Pharmacogenetics (see Chapter 43) is the name given to the science relating the effects of genes on our response to drugs, and is a rapidly expanding field within molecular genetics that is likely to become established within the repertoire of diagnostic clinical biochemistry laboratories in the future.

Monitoring the progression of disease

Although biochemical data alone may be of limited use in diagnosis, serial measurements can be of considerable value in monitoring the course of a disease or its response to treatment. The more closely the variable being measured relates to the underlying pathological process or functional abnormality, the better it will be for this purpose.

However, the reason for a change in a biochemical variable is not always the most obvious (or hoped for), so that even if an observed change is as expected or desired, the result should be interpreted with care. For example, a decrease in urine protein excretion in a patient with glomerular disease may indicate resolution of the underlying condition, but it could also be a result of deterioration leading to a decrease in the glomerular filtration rate. Biochemical data must always be interpreted in the light of clinical assessment and the results of other relevant investigations, not in isolation. Nevertheless, intervention may sometimes be appropriate on the basis of a

biochemical change alone, if this has been shown reliably to predict a significant clinical change, for example in hyperkalaemia in a patient with renal failure.

When serial biochemical measurements are used to follow the response to treatment, the failure of an expected change to occur may suggest that the treatment is inadequate or inappropriate, or even that the diagnosis is incorrect. In therapeutic drug monitoring (TDM, see Chapter 39), biochemical measurements may actually indicate a possible cause for non-response to treatment.

Biochemical investigations can also be used to detect the development of complications of diseases or their treatment before these become obvious clinically, and thus allow appropriate action to be taken before there is any clinical deterioration. They may even be used to prevent complications: for certain drugs, TDM allows presymptomatic detection of potentially toxic concentrations of the drug.

Screening

Screening for disease implies attempting to detect disease before it becomes manifest through the development of a clinical disturbance. Inherent in the concept of screening is that appropriate management of subclinical disease is of potential benefit to the patient. Screening can involve clinical assessment and laboratory and other investigations. For some conditions (particularly inherited metabolic diseases), screening may involve a single biochemical test. But the term is also used in relation to the performance of a range of biochemical tests (often combined with other types of investigation) in healthy people, in an attempt to detect any of a number of conditions, in the belief that a set of 'normal' results – that is, results within the appropriate reference limits – excludes these conditions. As will be seen in Chapter 2, considerable care is required both in the devising of screening tests and in their interpretation. While a set of 'normal' results may appear reassuring and may, indeed, exclude the presence of certain diseases, it may also convey a false impression and even delay the diagnosis of early disease. Because of the way in which reference ranges are defined, the more tests that are performed, the more likely it is that an 'abnormal' result (i.e. outside the reference limits) will be generated that is not related to the presence of disease.

A positive result in a screening test on its own should not usually be regarded as being diagnostic. When the prevalence of a condition in the population being screened is low, the predictive value of a positive result is often lower than is generally supposed. A positive result in a screening test must always be confirmed by further investigation. The use of direct methods (e.g. oligonucleotide probes, see Chapter 43) to detect mutations in DNA is an exception to this. Properly conducted, they are definitive with regard to the detection of mutations, although not necessarily for the development of disease.

Screening may be applied to a population, to groups sharing a common characteristic within a population or to individuals. According to the nature of the condition in question, screening may be carried out antenatally, shortly after birth, during childhood or during adult life. The strategy adopted will depend on the risk of the condition, the probability of its presence, the availability

of suitable screening tests and, inevitably, the cost. The latter includes particularly the economic cost of the programme but also the personal cost to individuals, for example those who test 'false positive' (individuals identified by the programme but on further investigation found not to have the condition in question) and, with inherited diseases, the relatives of individuals detected by the programme.

Population screening

Economic and logistic considerations preclude the screening of whole populations for disease, although it has been advocated, for example, that all adults (some suggest only males, others males and females) should be screened for hypercholesterolaemia. Although this would undoubtedly lead to the identification of a significant number of individuals at greatly increased risk of coronary heart disease because of severe but asymptomatic hypercholesterolaemia, such a programme, however desirable, would be very costly, and it has been argued that resources would be better devoted to measures to improve the general health of the population, by encouraging a healthy diet and lifestyle.

Selective screening

Selective biochemical screening for disease is already practised extensively in developed countries. The neonatal screening programmes for phenylketonuria, congenital hypothyroidism, sickle cell disease and cystic fibrosis are the best known examples. With the advent of tandem mass spectrometry, it is possible to screen for many more conditions such as medium chain fatty acid oxidation disorders, some organic acidaemias and the commonest form of congenital adrenal hyperplasia, using a tiny quantity of blood and at reasonable cost. These complement the thorough clinical screening of the newborn for conditions such as congenital cataract, imperforate anus etc.

Where a condition is particularly common in a defined group, screening may be appropriate, even though it would not be for the population at large. Antenatal screening for Tay–Sachs disease in Ashkenazi Jews is one example.

For hypercholesterolaemia, selective screening is a more practicable procedure than population screening. It can be applied to people in whom there is a high probability of hypercholesterolaemia being present, for example members of families in which there is a history of familial hypercholesterolaemia or premature heart disease. Such screening can also be directed towards people already at increased risk of coronary heart disease because, for example, they are smokers or have hypertension or type 2 diabetes, whose risk would be increased further by hypercholesterolaemia. Other examples of selective screening are discussed in the relevant chapters of this book.

Individual screening

Examples of individual screening include antenatal screening of a fetus for an inherited disease when a previous child of the parents has been found to have the condition

or when there is a strong family history of the condition. This has been practised for some time for certain inherited diseases, but the number for which it can be done is growing rapidly now that the mutations responsible for inherited diseases are becoming known. Although, undoubtedly, it will become possible to treat some of these conditions in utero, at present, antenatal screening is mainly aimed at detecting conditions with consequences so severe that it is considered appropriate to terminate the pregnancy if the genetic abnormality is present. Given this possible outcome, it is clearly essential that if the diagnosis is to rest only on the result of the screening test, this should provide accurate and unequivocal results.

Other uses of biochemical investigations

All the uses of biochemical investigations that have been discussed thus far are potentially of direct benefit to the patient. Other important uses include the provision of information for teaching, research and public health. Usually, this will relate to one of the categories discussed. Although such data may not be of immediate benefit to the patient, these areas are of immense potential benefit to the population, providing information fundamental to the advancement of knowledge. This use has ethical implications and is increasingly subject to scrutiny by bodies such as the network of research ethics committees in the UK. The use of biochemical investigations to assess organ function in potential transplant donors is an example of the use of investigations primarily for the benefit of other people. Data collection in specific disorders, such as for the UK Renal Registry for patients on dialysis, can help to improve the standard of care given to selected patient groups by comparing results achieved by different centres against pre-defined targets.

Extensive biochemical investigations are usually carried out during trials of drugs: these may be required as part of the assessment of a drug's efficacy but are also essential for the detection of possible toxicity.

Investigations may also be performed for the benefit of the doctor rather than the patient. Few doctors have not been guilty at some time, of requesting biochemical tests for reassurance. The supposition is that if the results of a range of test results are within reference limits, then the conditions in which abnormalities are known to occur cannot be present. As has been emphasized above, this supposition is erroneous and any reassurance may be unfounded. Biochemical investigations should be requested for one of the reasons discussed in the relevant chapter

of this book and not 'routinely'. Neither should junior medical staff be put under pressure to request unnecessary tests to placate their seniors.

It is regrettable that there is an increasingly perceived need for doctors to carry out a comprehensive range of investigations in case of subsequent litigation. While this is understandable, it should not be necessary if investigations are requested and performed in response to the individual clinical circumstances. There will always be other investigations that could have been done, but no blame should be attached to a doctor who failed to carry one out if it was not indicated clinically, either on the basis of the known natural history of the disease or the predicted response to, and known complications of, treatment.

CONCLUSION

Biochemical data are used extensively in medicine, both in the management of patients and in research. But before an investigation is requested, the rationale for testing should always be considered. Automated analysers can perform many tests at a very low cost in relation to the total expenditure on healthcare, but the cost is not negligible. There may also be a cost to the patient. Repeated venepunctures to obtain blood for 'routine' tests are at best a nuisance, and at worst can, particularly in small children, cause a significant fall in the haematocrit. The laboratory handbook at one hospital of the authors' acquaintances used to contain the following advice to junior medical staff: 'If you need advice or time to think, ask for it; do not ask for a full blood count and measurement of "urea and electrolytes".' In common with other investigations, biochemical investigations should be requested to answer specific questions; if there is no question, the result cannot provide an answer.

Further reading

Asher R. Richard Asher talking sense. A selection of his papers edited by Sir Francis Avery Jones. London: Pitman Medical; 1972.

A collection of essays by a clear thinking physician; his observations on the use of common sense in medicine, including the use of the laboratory.

Fraser CG. Interpretation of clinical chemistry laboratory data. Oxford: Blackwell Scientific; 1986.

A concise but comprehensive account of the uses of laboratory data, their acquisition and interpretation, relevant to this and the succeeding chapter, which should be required reading for clinical biochemists.

NCBI. One size does not fit all: the promise of pharmacogenomics, <http://www.auburn.edu/academic/classes/biol/3020/iActivities/CGAP2/Pharmacogenomics%20Factsheet.htm>; [Accessed 20.09.12].

A concise introduction to this topic.

Acquisition and interpretation of biochemical data

Helen Bruce • Marta Lapsley

CHAPTER OUTLINE

INTRODUCTION 6

THE TEST REQUEST 6

FACTORS AFFECTING TEST RESULTS 7

Preanalytical factors 7

Analytical factors 10

Postanalytical factors 12

INTERPRETATION OF RESULTS 13

Normal and abnormal 13

The meaning of normal 13

Reference values 14

Comparison of observed results with reference limits 15

Comparison of results with previous values 15

THE PREDICTIVE VALUE OF TESTS 16

Introduction 16

Prevalence and predictive value 18

Practical applications of the predictive value model 19

Receiver operating characteristic curves 19

Likelihood ratios 19

CONCLUSION 20

INTRODUCTION

It was emphasized in Chapter 1 that all investigations in medicine should be performed to answer specific questions. Biochemical data obtained must be considered in relation to the reason for the request, and against the background of an understanding of the relevant normal physiological and biochemical mechanisms and the way in which these respond to disease. One of the objectives of any laboratory is to ensure these data are available in a timely manner, and generated efficiently. The achievement of this goal requires careful attention to every step in the process, from the ordering of the investigation, the collection of the specimen(s) required, their transport to the laboratory and analysis, to the delivery of a report to the clinician, appropriate action being taken and the effects of this action being assessed. Amongst these many steps, the interpretation of data by clinical biochemists adds considerably to the value of the data. The workload of most laboratories is so great that it would be impossible (as well as being unnecessary) to add such comments to all reports (e.g. where the results are clearly normal). Interpretative comments (which may be individual or rule-based) are more likely to be required for more unusual tests, and for requestors who have only limited experience of the investigation in question. Typical reports requiring more detailed

interpretation include those with borderline data, results that are not consistent with the clinical findings, apparently contradictory data and changes in biochemical variables during dynamic function tests.

THE TEST REQUEST

The first step in performing a biochemical investigation is for a request form to be completed, often electronically, which prompts the collection of the appropriate specimen(s) and instructs the laboratory on the investigations(s) to be performed. Depending upon the reason for the request, the expertise of the clinician and the practice of the laboratory, the request may simply be for one or more specified analyses on a body fluid; for a set (often referred to as a 'profile') of standard investigations (e.g. 'thyroid function tests'); for a more involved procedure such as a dynamic function test involving the collection of serial samples following a specific stimulus, or an open request to perform whatever assays are deemed appropriate by the laboratory staff to answer the question posed in the request. The majority of biochemical test requests fall into the first two categories.

The information that is required when a test is requested is summarized in [Table 2.1](#).

TABLE 2.1 Information required when requesting a biochemical investigation

Information	Reason
Patient's name, identifying (e.g. hospital) number ^a , date of birth and sex	Identification and (age, sex) interpretation of results
Home address for primary care patients	Address may be useful if grossly abnormal results are found
Return address (e.g. ward, clinic, surgery; telephone/pager number if urgent)	Delivery of report
Name of clinician (and telephone/pager number)	Liaison Audit Billing
Clinical details (including drug treatment)	Justification of request Audit Interpretation Selection of appropriate assays Choice of analytical method (to avoid drug interference)
Test(s) requested	Instruction to analyst
Sample(s) required	Instruction to phlebotomist
Date (and time ^b if appropriate)	Identification Interpretation (with timed/sequential requests)

^aIn the UK, it is mandatory to supply the National Health Service Number.

^bIt is good practice also to record the time at which the specimen and request is received in the laboratory.

FACTORS AFFECTING TEST RESULTS

The generation of biochemical data is potentially subject to error at every stage in the process. It is essential that the sources of error are identified and understood, so that their effects can be minimized.

The sources of errors in biochemical tests are conventionally described in three categories:

- preanalytical: that is, either outside or within the laboratory, but before the analysis is performed
- analytical: these may be random (e.g. due to the presence of an interfering substance in the specimen) or systematic (e.g. because of a bias in the method)
- postanalytical: that is, occurring during data processing or transmission, or in relation to the interpretation of the data.

Preanalytical factors

Preanalytical factors may appear to be beyond the remit of clinical biochemists, but accreditation bodies are increasingly expecting laboratories to take responsibility for all aspects of testing. Laboratories should ensure that clinicians requesting investigations and the staff responsible for collecting the specimens understand the problems that can arise, so that specimens are collected and transported appropriately.

Preanalytical factors fall into two categories: those that relate to the specimen obtained for analysis (technical factors) and those that relate directly to the patient (biological factors).

Technical factors

These include:

- correct identification of the patient
- appropriate preparation of the patient where necessary
- collection of the specimen into an appropriate container with the correct anticoagulant or preservative
- accurate labelling of the specimen container after the sample has been collected (not before, as this carries a higher risk of a specimen being put into a container bearing another patient's name). Primary labelling of the sample with a barcode at source reduces the risk of mislabelling in the laboratory
- rapid and secure transport to the laboratory. Some specimens need to be transported under special conditions, for example arterial blood for 'blood gases' in a sealed syringe in an ice-water mixture; requests for blood, urine or faecal porphyrins must be protected from light.

Care must be taken during specimen collection to avoid contamination (e.g. with fluid from a drip), haemolysis of blood or haemoconcentration (due to prolonged application of a tourniquet). Appropriate precautions are also required during the collection and transport of urine, faeces, spinal fluid or tissue. Specimens known to be infective (e.g. from patients carrying the hepatitis B or HIV viruses) are occasionally handled specially, but it is good practice to handle all specimens as if they were potentially hazardous.

On receipt in the laboratory, the patient's name and other identifiers on the specimen must be checked against these details on the request form, whether paper or electronic. The specimen and form should then be labelled with the same unique number. As electronic requesting becomes more common, the majority of specimens are now being labelled with a unique barcode at source. Secondary containers (e.g. aliquots from the primary tube) should also be identified with patient details and the same unique number as the primary container. Automated preanalytical robotics and track systems can minimize the number of manual interventions required during sample handling, thereby minimizing risk of errors. All these processes are facilitated if computer systems in hospitals or surgeries have electronic links to the laboratory information system.

Laboratories should have written protocols (standard operating procedures) for the receipt and handling of all specimens to ensure the positive identification of specimens throughout the analytical process.

Biological factors

Numerous factors directly related to the patient can influence biochemical variables, in addition to pathological processes. They can conveniently be divided into endogenous factors, intrinsic to the patient, and exogenous factors, which are imposed by the patient's circumstances. They are summarized in [Table 2.2](#).

In addition, all biochemical variables show some intrinsic variation, tending to vary randomly around the typical value for the individual.

TABLE 2.2 Some biological factors affecting biochemical variables

Factor	Example
Endogenous	
Age	Cholesterol Alkaline phosphatase Urate
Sex	Gonadotrophins Gonadal steroids
Body mass	Triglycerides
Exogenous	
Time	Cortisol (daily) Gonadotrophins (in women, catamenial) 25-Hydroxyvitamin D (seasonal)
Stress	Cortisol Prolactin Catecholamines
Posture	Renin Aldosterone Proteins
Food intake	Glucose Triglycerides
Drugs	see Table 2.3 and text

Endogenous factors

Age. The reference values (mentioned in more detail later in the chapter) for many biochemical variables do not vary significantly with age during adult life. Some, however, are different during childhood, particularly in the neonatal period. A well-known example is plasma alkaline phosphatase activity, which is higher in children, particularly during the pubertal growth spurt, than in adults. Plasma cholesterol concentrations tend to increase with age, but may fall slightly over the age of 70; plasma urate concentrations tend to rise with age. Given that renal function tends to decline with age, it might be anticipated that mean plasma creatinine concentrations would rise with age, but the tendency for loss of muscle bulk in the elderly has a balancing effect. Other age-related changes are discussed elsewhere in this book.

Sex. Apart from the obvious differences in plasma gonadal hormone concentrations between adult men and women, other analytes demonstrate sex-related differences in concentration, often because their metabolism is influenced by gonadal hormones. Thus, total cholesterol concentrations tend to be higher in healthy men than in women until the menopause, after which concentrations in women tend to rise. In general, sex-related differences in biochemical variables are less between boys and girls prepubertally, while the differences between adult males and females decrease after the menopause.

When age and sex are important determinants of the level of a biochemical variable, measurements in patients should be considered in relation to age- and sex-related reference values, if valid conclusions are to be drawn.

Changes in many biochemical variables occur during pregnancy and, where necessary, measurements must be compared with reference values appropriate to the stage of gestation. The clinical biochemistry of pregnancy is considered in detail in Chapter 22.

Ethnic origin. Plasma creatine kinase activity tends to be higher in people of sub-Saharan African descent

TABLE 2.3 Some examples of in vivo effects of drugs on biochemical variables

Drugs	Effect	Mechanism
Glucocorticoids	Decreased plasma cortisol	Suppression of ACTH secretion
Thiazide diuretics	Decreased plasma potassium	Increased renal potassium excretion
Oestrogens	Increased plasma total thyroxine	Increased synthesis of binding proteins
Phenytoin, phenobarbitone, alcohol	Increased plasma γ -glutamyltransferase	Increased enzyme synthesis (enzyme induction)

than Caucasians (typically up to three times the upper reference limit; people of southern Asian origin may have intermediate values), but otherwise, there are no significant differences in the typical values for most biochemical variables between individuals of different ethnic origin living in the same region.

Body mass. Obese individuals tend to have higher plasma insulin and triglyceride concentrations than lean individuals, with an increased risk of developing type 2 diabetes and cardiovascular disease. Creatinine production is related to muscle bulk and plasma concentration may be above the usual reference range in a muscular individual, despite having a normal glomerular filtration rate. Twenty-four hour urinary excretion of many substances is greater in people of higher body mass. On the whole, however, body mass has little effect on the *concentrations* of substances in body fluids, although, of course, it is an important determinant of the total *quantities* of many substances in the body.

Exogenous factors. Many exogenous factors can have profound influences on the concentrations of biochemical variables even in healthy individuals. They include the time of day, stress, posture, fasting status, drugs, exercise and concurrent illness (see Tables 2.2 and 2.3).

Time-dependent changes. Rhythmic changes occur in many physiological functions and are reflected in changes in the levels of biochemical variables with time. The time base may be diurnal (related to the time of day, but usually nycthemeral, i.e. related to the sleep-wake cycle), catamenial (relating to the menstrual cycle) or seasonal. In addition, some hormones are secreted in sporadic bursts (e.g. growth hormone); when this occurs, it may be helpful to collect several blood samples over a short period of time and to base clinical decisions on the mean value.

The best known analyte having a diurnal variation in concentration is cortisol. Its concentration is at a nadir at about midnight, rises rapidly to reach a peak at 08.00–09.00h and then declines throughout the day. Observed values must be compared with reference values for specific times, with sampling at expected peak or trough times being the most informative. Other analytes

showing a diurnal rhythm (but to lesser extents) include thyroid stimulating hormone, testosterone and prolactin.

Some analytes show regular variations with a different time base. In women, during the reproductive years, the menstrual cycle is associated with regular changes in the concentrations of gonadotrophins, oestrogens and progesterone. Measurements made for diagnostic purposes must be made at the appropriate time in the cycle: for example, an increase in plasma progesterone concentration seven days before the onset of the next menstrual period is due is taken as an indication that ovulation has occurred.

Plasma 25-Hydroxycholecalciferol concentration varies with the season, being higher in summer than in winter.

Stress. Concern for the patient dictates that stress should be minimized at all times, but this is particularly important in relation to blood sampling for those analytes with concentrations that are responsive to (usually by increasing) stress. Pituitary and adrenal hormones are particularly affected. Thus plasma adrenocorticotrophic hormone (ACTH), cortisol, prolactin, growth hormone and catecholamine concentrations all rise in response to stress. Indeed, this effect is utilized in investigations of pituitary function. However, avoidance of stress is vital when collecting specimens for the measurement of these hormones under other circumstances.

Posture. Posture has a significant effect on a wide range of analytes. The best known of these are plasma renin activity and aldosterone concentration. Both are higher in the standing than the recumbent (or even sitting) position, particularly shortly after the change in posture, as a result of a decrease in renal blood flow.

The effect of posture on certain other analytes is less well appreciated. When people are upright, there is a greater tendency for fluid to move from the vascular to the interstitial compartment than when they are recumbent. Small molecules and ions in solution move with water, but macromolecules and smaller moieties bound to them do not. As a result, the concentrations of proteins, including lipoproteins, and of protein-bound substances, for example thyroid and other hormones, calcium and iron, tend to be approximately 10% higher when an individual is in the upright position than when recumbent. Intermediate values occur when sitting. This effect may be relevant when values obtained from individuals when they are outpatients are compared with values obtained when they are inpatients.

Food intake. The concentrations of many analytes vary in relation to food intake. Frequently encountered examples include glucose, triglycerides and insulin, the plasma concentrations of which all increase following a meal, so that they should usually be measured in the fasting state, unless the effect of recent intake (as in the glucose tolerance test) is being examined. Some specific dietary constituents can affect biochemical variables; consumption of red meat in the hours before venepuncture may increase plasma creatinine concentrations by as much as 30% from fasting values. A protein-rich meal results in increased urea synthesis and increases plasma urea concentration. Long-term dietary habits can also significantly affect biochemical variables (e.g. cholesterol).

The urinary excretion of many substances is highly dependent on their intake and, although some laboratories publish reference ranges for them, they are very wide. In assessing the significance of the excretion of substances in the urine, it is important to consider their intake and thus their expected excretion if renal function is normal. Thus, for example, a low urinary sodium excretion is a normal response to sodium depletion.

Drugs. Drugs, whether taken for therapeutic, social or other purposes, can have profound effects on the results of biochemical tests. These can be due to interactions occurring both in vivo and in vitro.

In vivo interactions occur more frequently. They can be due to direct or indirect actions on physiological processes or to pathological actions. There are numerous examples in each category. Some of the better known physiological interactions are indicated in Table 2.3. Others are discussed elsewhere in this book.

Pathological consequences of drug action in vivo can be idiosyncratic (unpredictable) or dose related. A measurable biochemical change may be the first indication of a harmful effect of a drug, and biochemical tests are widely used to provide an early indication of possible harmful effects of drugs, both for those in established use (e.g. measurement of serum creatine kinase activity in patients being treated with statins) and in drug trials.

Drugs can also interfere with analyses in vitro (this is strictly an analytical factor, but is mentioned here for completeness), for example by inhibiting the generation of a signal or by cross-reacting with the analyte in question and giving a spuriously high signal. This field is well documented, but the introduction of new drugs means that new examples of this source of error are continually being described.

Other factors. Exercise can cause a transient increase in plasma potassium concentration and creatine kinase activity; the latter used to be a potential confounding factor in the diagnosis of myocardial infarction in a patient who had developed chest pain during exercise before cardiac-specific troponin measurement was introduced. Even uncomplicated surgery may cause an increase in creatine kinase as a result of muscle damage, and tissue damage during surgery can also cause transient hyperkalaemia. Major surgery and severe illness can elicit the 'metabolic response to trauma' (see Chapter 20), which can lead to changes in many biochemical variables.

Intrinsic biological variation. The contributions to biological variation discussed thus far are predictable, and either preventable or can be allowed for in interpreting test results. However, levels of analytes also show random variation around their homeostatic set points. This variation contributes to the overall imprecision of measurements and must also be taken into account in the interpretation of test results. It is also relevant to the setting of goals for analytical precision (see below).

Intrinsic biological variation can be measured by collecting a series of specimens from a small group of comparable individuals over a period of time (typically several weeks) under conditions such that analytical imprecision is minimized. The specimens should be handled identically and analysed in duplicate, using an identical technique (e.g. the same instrument, operator, calibrators and

reagents). This may either be done at the time the specimens are collected or in a single batch. If the latter approach is adopted, the specimens must be stored in such a way that degradation of the analyte is prevented (e.g. by freezing at a sufficiently low temperature). Using a single batch procedure is preferable as it eliminates between-run analytical variation.

If the specimens are analysed in a single batch, the analytical imprecision (see below) can be calculated from the differences in the duplicate analyses and is given by:

$$SD_A = \sqrt{\frac{\sum (\text{differences})^2}{2n}}$$

where SD is the standard deviation and n is the number of pairs of data. If, however, specimens are analysed at the time they are taken, the analytical imprecision must be calculated from replicate analyses of quality control samples.

The standard deviation of a single set of data from each individual is then calculated after excluding any outliers ('fliers') using an appropriate statistical test. This standard deviation will encompass both analytical variation and the intraindividual variation (SD_I), such that:

$$SD = \sqrt{SD_A^2 + SD_I^2}$$

Since SD_A is known, the intraindividual variation can then be calculated. It is also possible to calculate the interindividual variation (SD_G , due to the difference in the individual homeostatic set points for the analyte between each member of the group) by calculating the SD for all single sets of data for all subjects in the group, since this SD is given by:

$$SD = \sqrt{SD_A^2 + SD_I^2 + SD_G^2}$$

The calculations required can be performed using a nested analysis of variance technique as provided in many statistical software packages. An alternative approach is to calculate coefficients of variation (CV) (given by $(SD \times 100)/\text{mean}$). Values for CV can be substituted for SD in these formulae because the means are similar.

Some typical values of biological variation for frequently measured analytes are given in Table 2.4. (A reference to a more extensive list is provided in the Further reading section at the end of the chapter.) It should be noted that, while in most instances the interindividual variation is greater than the intraindividual variation, this is not always the case. When it is not, it means that the extent of natural variation around individuals' homeostatic set points is more than the range of variation between these set points. The relative sizes of intra- and interindividual variation have important consequences for the interpretation of analytical data in relation to reference ranges, as will be discussed later in the chapter.

Analytical factors

It is clearly essential that every effort is made to reduce the possibility of errors arising during analysis. A rigorous

quality assurance programme (discussed in Chapter 3), is required to ensure the quality of results. The production of reports and their transmission to the requesting clinician are discussed below (see Postanalytical factors).

Test methods and procedures must be selected to allow adequate performance. Some of the important factors that impinge on performance will now be considered. The question of what is adequate performance will also be discussed. It should be appreciated that deciding on the standards for performance of individual assays requires an appreciation of all the factors involved: efforts to reduce specimen volume (an important consideration in paediatric, and particularly neonatal, practice) may affect accuracy and precision; these may also be influenced by a requirement to achieve a very rapid turnaround time. Cost, instrumentation and staff skill mix are other important factors.

Analytical range. The test selected must be capable of measuring the analyte in question over the whole range of concentrations that may be expected to occur. Measuring large quantities of an analyte is usually easier than measuring small quantities (though dilution of the sample may be necessary to bring the concentration within the range of acceptable performance, i.e. to avoid nonlinearity of signal to concentration). Note, however, that if dilutions are to be employed, their validity must be established by appropriate studies to demonstrate that dilution, per se, does not affect the relationship between analyte concentration and the signal generated. It is not always possible to design an assay that has a low enough detection limit (i.e. the lowest level of the analyte that can be reliably distinguished from zero, usually taken as the mean signal generated by zero standards plus two standard deviations); this has been a particular problem with the analysis of certain hormones, the plasma concentrations of which are in the nano- or even picomolar range, but is becoming less so with the introduction of more sensitive assay techniques.

For some analytes, the need to be able to measure concentrations over very wide ranges can pose technical problems, for example the 'high dose hook effect' with immunoassays for prolactin and human chorionic gonadotrophin. This term refers to the decrease in signal that occurs at high concentrations as a result of binding sites on the capture and labelled antibodies being occupied by separate molecules of the analyte rather than these causing crosslinking between them (the usual situation when the antibodies are in excess). For C-reactive protein (CRP) in serum, the requirement to be able to measure concentrations accurately in two different ranges ('high sensitivity' CRP as a cardiovascular risk marker or marker of neonatal sepsis and CRP as an inflammatory marker) has led to the introduction of separate assays for each of the two ranges of concentration.

Accuracy and bias. Accuracy reflects the ability of an assay to produce a result that reflects the true value. Inaccuracy is the numerical difference between the mean of a set of replicate measurements and the true value. The term 'bias' is often preferred to 'inaccuracy' in laboratory medicine as it implies a persistent characteristic rather than an

TABLE 2.4 Biological and analytical variation and consequent desirable analytical performance parameters for some frequently measured constituents of plasma expressed as coefficients of variation (CV=(SD×100)/mean)

Analyte	Intra-individual variation (%)	Inter-individual variation (%)	Typical analytical imprecision (%)	Desirable analytical precision (%)	Desirable analytical bias (%)
Sodium	0.7	1.0	1.1	0.4	0.3
Potassium	4.8	5.6	1.5	2.4	1.8
Calcium (total)	1.9	2.8	2.6	1.0	0.8
Magnesium	3.6	6.4	4.5	1.8	1.8
Phosphate	8.5	9.4	3.0	4.3	3.2
Bicarbonate	4.8	5.3	7.4	2.4	1.8
Urea	12.3	18.3	4.6	6.2	5.5
Creatinine	6.0	14.7	7.9	3.0	4.0
Urate	9.0	17.6	3.3	4.5	4.9
Glucose	4.5	5.8	2.5	2.3	1.8
Bilirubin (total)	23.8	39.0	6.0	11.9	11.4
Cholesterol (total)	5.4	15.2	3.0	2.7	4.0
Triglycerides	20.9	37.2	4.1	10.5	10.7
Total protein	2.7	4.0	3.7	1.4	1.2
Albumin	3.1	4.2	3.4	1.6	1.3
Alkaline phosphatase	6.4	24.8	4.7	3.2	6.4
Alanine aminotransferase	18.0	42.0	4.7	9.0	11.4
γ-Glutamyl transferase	13.8	41.0	4.2	6.9	10.8
Creatine kinase	22.8	40.0	4.2	11.4	11.5

Data from various sources, including UKNEQAS (UK National External Quality Assurance Scheme) website and Ricos et al. on Westgard QC website (see Further reading).

occasional effect. It is due to systematic error (*cf.* precision, below) and can be positive (the result is higher than the true result) or negative (lower than the true result), and constant (being the same absolute value throughout the analytical range) or proportional (having the same *relative* value to the true result).

Ideally, analyses should be without bias, but, provided that its extent and characteristics are known, the predictability of bias makes it less of a problem in the interpretation of laboratory data than imprecision.

Precision. Precision is a reflection of reproducibility. Imprecision is defined as the standard deviation for a series of replicate analyses (i.e. made on the same sample, by the same method, under identical conditions). For many analytes it is very low, giving coefficients of variation (CV, defined as (SD×100)/mean value) of 1% or less. For others it is higher, giving CVs as high as 5%. The median imprecisions for some serum analytes, as achieved by a group of laboratories in the UK, are given in Table 2.4.

Imprecision in analysis can never be entirely eliminated. Such analytical variation is an important factor to be borne in mind when interpreting the results of laboratory tests, as discussed later.

Analytical variation or imprecision can be measured by performing repeated analyses on a single sample and calculating the standard deviation of the results, or by doing duplicate measurements on a series of samples as described in the section on biological variation, above. Imprecision tends to be highest when the signal generated by an assay (e.g. a colour change or electric potential) is low. To be of value in interpreting results, imprecision should be determined over a range of concentrations and particularly in the region of critical values (decision levels), where a

small change in a result can lead to a significant change in management of the patient.

Specificity and interference. Specificity is the ability of an assay to measure only the analyte it purports to measure. Most assays are highly specific, but exceptions do occur, both owing to endogenous substances and to drugs that can give a signal in the assay and cause the concentration/activity to appear higher or lower than it is. A detailed discussion of this topic is beyond the scope of this chapter. Some important examples are mentioned elsewhere in this book and further information is available from the bibliography. Interference is a related, but separate problem, in that a substance alters the signal given by the analyte in question, but does not generate a signal itself.

Practicalities: what is desirable performance? In their enthusiasm to perfect analytical methods, clinical biochemists must be careful not to lose sight of the application of their work. A result is of no value if its bias or imprecision render it dangerously unreliable; neither is it of value if the method is practically too complicated or too expensive to be useful, or if it takes so long to perform that the need for the result has passed by the time it becomes available.

What is required of a test is that it should be capable of providing a result that will answer the question that is posed. To give a simple example, if the requirement is to confirm that a patient with the typical clinical features of diabetic ketoacidosis does indeed have this condition, what is needed is a rapid test that will confirm the presence of hyperglycaemia. In this context, it is of little relevance whether the blood glucose concentration is 30 or 35 mmol/L. But if a glucose tolerance test is required to establish (or refute) a diagnosis of diabetes because the

clinical features are equivocal, time is of little concern, but the precision and accuracy are of paramount importance. Imprecision or bias may lead to the diagnosis being missed or a patient being classified as having diabetes and managed as if he had the condition when he does not.

In practice, it is rarely necessary to sacrifice accuracy and precision for the sake of a result being available rapidly. Providing that they are used competently, and in accordance with standard operating procedures, many rapid methods, including point-of-care (near-patient) testing instruments, using 'dry chemistry' techniques, are intrinsically as reliable as traditional laboratory-based procedures. But, although clinical biochemists rightly strive to increase the accuracy and precision of assays, there must come a point where further efforts may not result in significant improvements in the reliability of data or their value when applied to patient care. The question: 'What is desirable performance?' needs examination in some detail.

Analytical goals. Numerous strategies have been applied to the setting of desirable standards for the performance characteristics of laboratory tests or 'analytical goals'. These have been based, for example, on the opinions of clinical users, of laboratory-based experts, on imprecision data, on arbitrary fractions of the reference range etc. The most widely accepted strategy for defining analytical goals is based on data on analytical imprecision in the assay and biological variation for the analyte. The *desirable* goal is that analytical imprecision should be less than or equal to half the intraindividual biological variation.

Expressing imprecision as coefficients of variation (CV), the goal is thus: $CV_A \leq 0.5 \times CV_I$, where CV_A is the CV for analytical imprecision and CV_I is the CV for intraindividual variation. Since the overall imprecision is equal to the square root of the sum of the squares of individual imprecisions, i.e.:

$$CV_{TOTAL} = \sqrt{CV_A^2 + CV_I^2}$$

it follows that if the goal is achieved, $CV_{TOTAL} \sim 1.12 \times CV_I$. This means that if the goal is achieved, analytical imprecision will contribute only about 12% to the overall imprecision of the result.

This concept has been extended to encompass goals for *optimum* performance ($CV_A \leq 0.25 \times CV_I$) and *minimum* performance ($CV_A \leq 0.75 \times CV_I$). Clearly, the less the analytical precision, the less the overall precision, but at any level of biological imprecision, equal increments in analytical precision (that is, lower imprecision) will produce relatively smaller increments in the overall quality of the result. Striving for analytical perfection may, therefore, be an inappropriate use of resources, which does not provide additional clinical value in decision-making. Some examples of analytical goals are given in Table 2.4. The relevance of both analytical and biological variation to the interpretation of results must not be understated. High precision and accuracy must be obtained over the whole range of concentrations that are likely to be encountered, but particularly over the range(s) critical to decision-making.

This brief discussion of analytical goals has centred on imprecision. Bias may appear less important, since most results on most patients will be compared with their own previous values or with reference ranges obtained using the same method. But patients may move from one healthcare setting to another and analyses may be performed by different methods in a variety of settings, so that accuracy (or lack of bias) to ensure comparability of results is also important. This is particularly the case when diagnoses and management decisions may be based on consensus opinions, for example the World Health Organization criteria for the diagnosis of diabetes, targets for glycated haemoglobin in the management of diabetes or the recommendations of various expert bodies for the management of hyperlipidaemias. An alternative approach to minimizing differences between methods is to develop nationally agreed guidelines or standards. In the UK, the introduction of the modified Modification of Diet in Renal Disease (MDRD) formula for calculating estimated glomerular filtration rate (eGFR) that is traceable back to the reference method for measuring creatinine (see Chapter 7), is an example of such standardization.

Desirable standards for bias have also been set and can be combined with those for precision to produce standards for total allowable error. For analytical bias (B_A), the desirable goal is given by:

$$B_A = < 0.25 \sqrt{(CV_I^2 + CV_G^2)}$$

This can be combined with the analytical goal for imprecision to produce the 'total allowable error' (TE_a) given by:

$$TE_a = < 0.25 \sqrt{(CV_I^2 + CV_G^2)} + 1.65(0.5 \times CV_I)$$

The interested reader is referred to the detailed discussion of this topic by Fraser (see Further reading, below). In practice, these concepts are less widely applied to assay performance than perhaps they should be, but ideally, their value should be assessed for individual analytes on the basis of their influence on clinical outcome.

Postanalytical factors

Errors can still arise even after an analysis has been performed, for example if calculations are required or if results have to be transferred manually, either directly from an analyser to a report or entered into a computer database. Even computers are not immune to error.

An error in the transmission of the report to the responsible clinician can also lead to an incorrect result being used for the management of the patient. Transcription errors can arise if results are telephoned, although it may be essential to do this if results are required urgently. Strict procedures should be followed when telephoning reports, both in the laboratory and by the person receiving the report. Although time-consuming, it is advisable for the person receiving the report to read it back to the person telephoning, to ensure that it has been recorded correctly.

Increasingly, data transfer is direct from the laboratory to the electronic patient record (EPR) and paper reports may not be printed routinely. Where paper reports are still used they must be transferred to the patient's notes. Even at this stage (especially if there are patients with the same or similar names in the same area), care must be taken; misfiled reports are at best wasted and at worst, if acted on as if they referred to another patient, positively dangerous.

The final stage in the process is for the requesting clinician to take action on the basis of the report. The report must be interpreted correctly in relation to the clinical situation and the question that prompted the request. This interpretation may be provided by a member of the laboratory staff, by the requesting clinician or may be a matter for discussion between them. However, errors can arise even at this stage, leading to an inappropriate clinical decision. The interpretation of data is discussed in the next section. Requesting laboratory investigations can appear to be a simple process, but it should by now be apparent that there are many potential sources of error. All laboratories should engage in a programme of audit, whereby action is taken to maintain and improve performance on a continuing basis, which will be discussed in Chapter 3.

INTERPRETATION OF RESULTS

Normal and abnormal

In interpreting the results of a biochemical investigation, one of two questions is likely to be posed: 'Is the result normal?' or, if it has been performed before, 'Has the result changed significantly?'. Arguably, a more relevant question to ask is, 'What does this result add to my knowledge that will allow me to manage this clinical situation better?'. The importance of considering this broader question should become apparent from a consideration of the problems associated with answering the first two, narrower, questions.

When an investigation is performed on an individual for the first time, the result must be assessed against a reference point. Usually, it is assessed specifically against what is expected in a healthy individual, although it might be more relevant in a symptomatic individual to assess it against what is expected in a comparable patient with similar clinical features. The range of values expected in healthy individuals has often been termed the 'normal range', but for various reasons, the term 'reference range' (strictly, 'reference interval') is now preferred.

The meaning of normal

Normal ranges have traditionally been defined on the basis of measurements of the analyte in question in a sufficiently large sample of individuals from an appropriate (in terms of, e.g. age, sex or ethnicity) healthy population. For data having a Gaussian distribution, the normal range is defined as the range of values lying between the limits specified by two standard deviations below the mean and two standard deviations above (Fig. 2.1). This

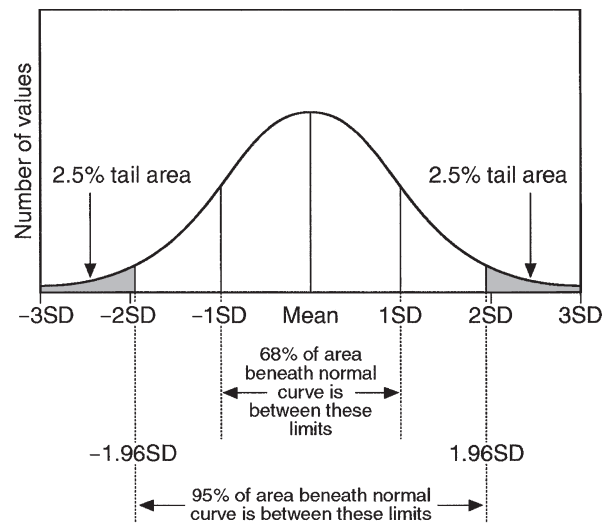


FIGURE 2.1 ■ The normal (Gaussian) distribution. The normal range encompasses values falling between two standard deviations above and below the mean and includes approximately 95% of all the values.

range encompasses 95% of the values found in the sample. The implication is that the great majority of healthy people will have a value for the analyte within this range.

'Normal' is a term used in statistics to describe a Gaussian distribution. Although the term 'normal range' is statistically valid only when the distribution is Gaussian, many analytes frequently measured in the laboratory have a skewed distribution – most frequently with a skew (tail) towards higher values. Examples include plasma alkaline phosphatase activity and the concentrations of triglycerides and bilirubin. A further drawback to the use of the term 'normal range' is that the word 'normal' has several different meanings in addition to its use in statistics. Given that the normal range, as defined above, encompasses only 95% of values characteristic of the chosen sample, 5% of individuals are bound to have values outside that range and might thus be considered 'abnormal'. This is clearly absurd in relation to the more colloquial meanings of the word, for example 'usual', 'acceptable', 'typical' or 'healthy'.

Often there is not a definite cut-off associated with an analyte, but a continuum of increasing risk with increasing concentration. An excellent example of this is the risk of coronary heart disease associated with increasing plasma cholesterol concentration, which extends below even the value of 5.0 mmol/L, often regarded as being an appropriate target for healthy (low risk) individuals. Indeed, although the curve of risk against cholesterol concentration becomes very shallow at low concentrations, it does not entirely flatten out (Fig. 2.2): that is, there is a continuum of increasing coronary risk associated with higher plasma cholesterol concentrations even in the lower part of the range of values typically found in adults in the UK. In this context, then, the term 'normal range' is potentially misleading: it is more relevant to define target values, which will depend on the overall level of coronary risk.

Furthermore, even in patients with established disease, results may fall within the 'normal range'. For most analytes, there is an overlap between the values usually

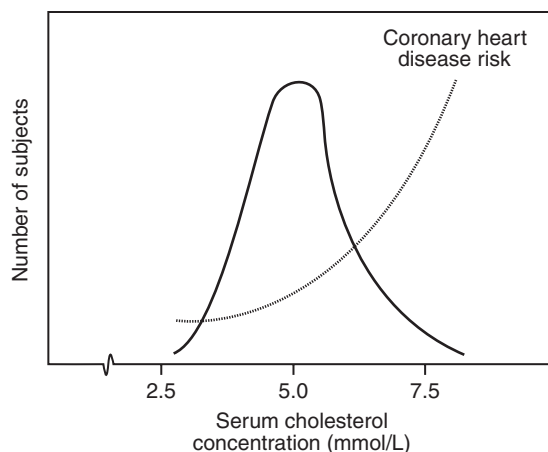


FIGURE 22 ■ The distribution of serum cholesterol concentration in healthy subjects and risk of coronary heart disease. More than one-third of the values exceed the recommended highest acceptable concentration (5.0 mmol/L) and are associated with significantly increased risk. Note that the distribution is non-Gaussian, being skewed to the right.

encountered in health and those encountered in disease (particularly if the disturbance is mild). Thus, it should not be assumed that all patients with values within the normal range for a particular analyte are free of the relevant disease, just as it has been indicated that (by definition) not all patients with values falling outside the range are necessarily abnormal in any other way.

Reference values

Such considerations led to the development, in the late 1960s, of the concept of reference values. The word 'reference' is free of the ambiguities associated with 'normal'. A reference value is defined as the value for an analyte obtained by measurement in an individual (reference individual) precisely selected using defined criteria, for example sex, age, state of health or other relevant characteristics.

The qualities required of a reference individual are only those specified: the term should not be taken as implying the possession of any quality (e.g. 'healthiness') that has not been specified. If measurements are made on a representative sample (reference sample) from a population of reference individuals (reference population) a distribution of values (a reference distribution) is obtained.

Reference limits can then be set and the range between these is defined as a reference interval. Typically, reference limits are set so that the interval encompasses values found in 95% of the reference samples, and this interval may then numerically be the same as the statistically defined normal range. Reference intervals are often colloquially termed 'reference ranges', but, strictly speaking, this is incorrect usage, since the true reference range encompasses the whole range of values derived from the reference population. The term 'reference range' is used in this book in accordance with common practice as a more familiar synonym for the term 'reference interval'.

Reference values can be obtained by carrying out measurements on a carefully selected and precisely defined

reference sample. If the distribution of values is Gaussian, the reference limits are the mean minus two standard deviations and the mean plus two standard deviations. When the distribution is non-Gaussian, the range can be calculated by ranking the values and deleting the lowest and highest 2.5%. Such a procedure has the considerable advantages that it requires no assumptions to be made about the characteristics of the distribution and does not require any transformation of the data to be made.

When the reference interval is derived from a sample of healthy individuals not known to be at increased risk of disease, the term effectively means 'values characteristically found in a group of healthy individuals and thus likely to encompass values found in other, comparable (with regard to age, sex etc.) healthy individuals'. It is sometimes called a 'health-associated reference interval'. Note, however, that health is not essential to the definition. Reference values could as well be established for analytes in disease as in health, although in practice this has not often been done (and, as indicated above, their ranges would almost always overlap with the corresponding health-related reference ranges).

Many strategies have been suggested for expressing observed values for biochemical variables in relation to the reference range: for example, the observed value can be expressed as a percentage of the mean of the reference interval or of the upper reference limit to give a 'centi-normalized unit'. Most such suggestions have not been adopted in clinical practice. For enzyme measurements, however, where the values obtained can be highly method-dependent and can, in disease, be orders of magnitude higher than reference values, it is sometimes helpful to express results as multiples of the upper reference limit. Measurements used in antenatal screening, e.g. α -fetoprotein and human chorionic gonadotrophin, are often reported as 'multiples of the median'.

An alternative suggestion has been to quote a 'reference change value', which looks at a patient's usual intra-individual variation and expresses the observed value as a percentage of the mean individual reference interval. The development of sophisticated laboratory information systems has enabled alternatives to the reference range to become a realistic possibility.

Problems with reference intervals

Even reference intervals have disadvantages. However derived, they encompass the values from only 95% of the reference population. Thus, 5% of reference individuals will have values outside the reference limits. If two analytes that vary independently are measured, the probability of one of them being outside its reference limits is $(1-0.95^2)$, that is 0.10. Of the reference individuals, 10% would be expected to have one of the measured values outside the reference limits. For 'n' analytes that vary independently, the probability is $(1-0.95^n)$, so that if 20 analytes were to be measured in an individual, there would be a 64% chance of one result being outside the reference limits. In practice, many tests are dependent on one another to some extent (e.g. total protein and albumin), so that this probability is reduced, but the calculation serves to emphasize that, even where values characteristic of

health and disease have a completely bimodal distribution (i.e. non-overlapping ranges, something that happens very rarely in practice), by no means all observed results outside the health-related reference limits would be associated with disease. On the other hand, common sense dictates that the further away from the reference limit an observed result is, the more likely it is to be pathological in origin.

Comparison of observed results with reference limits

The importance of comparing observed results with appropriate (e.g. age- and sex-specific) reference intervals has already been emphasized.

As has been discussed, the appreciation of analytical and intrinsic biological variation for each analyte is also important. The implications for the interpretation of observed data on patients are considerable. As an example, we can consider the measurement of creatinine in plasma, widely used as a test of renal function. Its (health-related) reference interval is of the order of 60–110 µmol/L, but this does not mean that this range of values will be observed over time for a single healthy individual. The intraindividual biological variation of creatinine concentration is around 4.6 µmol/L so that the reference interval for individuals is approximately 18 µmol/L. Thus, two healthy individuals from the reference population might have typical creatinine concentrations that ranged between, for example 60–78 µmol/L and 85–103 µmol/L, respectively. But, as will be discussed in the next section, a *change* in plasma creatinine concentration in an individual from 70 to 98 µmol/L would almost certainly signify a decrease in renal function, even though both results are well within the population-based reference interval.

When (as is usual) there is overlap, or even a continuum, of values of an analyte in health and disease, it may be necessary to define a cut-off value to determine if further action should be taken (e.g. initiating further investigations). This is particularly true of screening tests, when there may be no other information to take into account in deciding the significance of a result. The considerations on which the selection of such a cut-off value depends are discussed later in this chapter.

In summary, it is important to appreciate that the information that can be provided by a single biochemical measurement is limited. In general, the results of complementary or serial measurements are often of much greater value.

Comparison of results with previous values

The problem posed in the interpretation of a test result when the test has been performed previously, is a different one. The relevant questions are, ‘Has there been a change and, if so, is it clinically significant?’

In order to assess the relevance of any change to a pathological process, it is necessary to have a quantitative indication of the possible contribution of analytical and intrinsic biological variation to the observed change. The probability that a difference between two results (whether an increase or a decrease) is significant at the

95% level (i.e. that such a difference would only be expected to occur by chance alone on fewer than one in 20 occasions), requires that the difference is 2.8 times the variation (SD) of the test (for the basis of this calculation, see Fraser 2001, in Further reading, below). Because of the contribution of both analytical and biological variation, the standard deviation of the test is given by:

$$SD = \sqrt{SD_A^2 + SD_B^2}$$

where SD_A is the analytical variation and SD_B is the biological variation. One set of estimates of the value of $2.8 \times SD$, also known as the critical difference (more formally called the ‘reference change value’), for some frequently measured analytes, is given in Table 2.5. These values appear to differ little in health and disease and are valuable guides to clinical decision-making. Two caveats are important. First, the values quoted are mean values: individuals may show greater or lesser intrinsic biological variation. Second, analytical performance (and hence analytical variation) varies considerably between laboratories: ideally, laboratories should calculate their own values for critical differences, although in practice, this is seldom done.

It should be emphasized that, just because a change in the level of a biochemical variable exceeds the critical difference calculated from the known biological and analytical variation, it does not necessarily follow that it is clinically significant: that will depend on the precise clinical circumstances and needs to be assessed in relation to whatever clinical and other data are available. For example, restriction of dietary protein intake may cause a considerable reduction of plasma urea concentration

TABLE 2.5 Critical differences: changes required in the plasma level of some frequently measured analytes to be of significance in an individual

Analyte	Concentration	Change absolute	Change (%)
Sodium	140 mmol/L	5.1	4
Potassium	4.2 mmol/L	0.6	14
Calcium (total)	2.4 mmol/L	0.2	9
Magnesium	1.0 mmol/L	0.2	16
Phosphate	1.2 mmol/L	0.3	25
Bicarbonate	26 mmol/L	6.4	25
Urea	5.0 mmol/L	1.8	37
Creatinine	60 µmol/L	16.7	28
Urate	260 µmol/L	69.8	27
Glucose	4.6 mmol/L	0.7	14
Bilirubin (total)	10 µmol/L	6.9	69
Cholesterol	5.8 mmol/L	1.0	17
Triglycerides	1.2 mmol/L	0.7	60
Total protein	75 g/L	9.6	13
Albumin	40 g/L	5.2	13
Alkaline phosphatase	60 U/L	13.3	22
Alanine aminotransferase	20 U/L	10.4	52
γ-Glutamyltransferase	40 U/L	16.2	40
Creatine kinase	150 U/L	97.4	65

Data from various sources, including UKNEQAS (UK National External Quality Assurance Scheme) website and Ricos et al. on Westgard QC website (see Further reading).

in a patient with chronic renal disease, but this does not signify an improvement in renal function. The reverse is also true: the lack of a change does not necessarily imply that there is no cause for concern. For example, the haemoglobin concentration of blood may initially remain within health-related reference limits for a short period, even after severe acute bleeding, because erythrocytes and plasma have been lost together.

Another important point in relation to critical difference values, is a possible difference between laboratory staff and clinical staff in ascribing significance to an apparent change. The choice of a probability of 0.95 that two results are significantly different is arbitrary, although hallowed by practice. A probability of 0.99 is clearly more significant, but if two results differ with a probability of only 0.9, there is still a 9 in 10 chance that they *are* truly different. Decisions in medicine are still frequently made on much lower probabilities. It would be feasible to construct lists of critical differences at different levels of probabilities, but to publish such lists in laboratory handbooks would make them cumbersome and potentially confusing. However, laboratories should be able to provide guidance in this area to recently qualified clinicians and for newly introduced investigations.

In practice, critical values for difference used by clinicians are often empirical and not based on any formal assessment. This does not invalidate their use; indeed, they have often been validated informally by long experience and observation of outcomes. However, they fall short of the requirements of 'evidence-based clinical biochemistry' (discussed in Chapter 3).

Most instrument-based or laboratory-based management systems incorporate automatic procedures for identifying large (and potentially significant) changes in biochemical variables ('delta checks'). Despite every effort to reduce analytical variation, such changes may arise because of unusual analytical imprecision or random error. Gross analytical errors ('fliers') will often be obvious because they are greater than typically occurs clinically. Smaller errors may not be obvious. Deciding whether a change in a variable is likely to represent a real change in the patient requires checks on the identity of the specimen, internal quality control performance and the clinical details, if these are available on the request form or can be deduced from other information (the patient being located in the intensive therapy unit, for example), all of which the clinical biochemist will use in deciding what action to take. If the identity of the specimen is confirmed, the assay run was in control and the change is consistent with what is known about the patient, the result can be reported; if not, the original specimen may need to be re-analysed or a further specimen obtained from the patient.

As well as 'delta checks', laboratories should be able to identify action limits, such that unexpectedly high or low results that may require an immediate change of management (e.g. low or very high serum potassium concentrations) are checked and communicated rapidly to the requesting clinician. Specific values may be set to trigger further investigations on the basis of protocols agreed with the appropriate clinicians, for example magnesium if the corrected calcium concentration is low, or serum protein electrophoresis if the total globulin concentration is

increased. All these procedures are of potential value to the patient and add to the clinical usefulness of the results in question.

THE PREDICTIVE VALUE OF TESTS

Introduction

It will have become clear that the interpretation of biochemical data is bedevilled by the inevitable presence of analytical and intrinsic biological variation. It is still frequently assumed that a result within reference limits indicates freedom from disease or risk thereof, while a result outside reference limits at least requires that the patient be further investigated. What is lacking is any numerical indication of the *probability* that a particular result indicates the presence or absence of disease.

The predictive value concept, introduced into clinical biochemistry in the mid-1970s by Galen and Gambino, represents an attempt to address these shortcomings. Essential to the concept of predictive values is that disease or freedom from disease can be defined absolutely, that is, that there is a test which can be regarded as the 'gold standard'. For some conditions, this will be histological examination of tissue obtained at surgery or post-mortem, but it may be the eventual clinical outcome or some other more or less well-defined endpoint. It is against the gold standard that biochemical tests may be judged.

Definitions

If all the members of a population consisting of people with and without a particular disease are subjected to a particular test, the result for each of them will fall into one of four categories:

- true positives (TP) – individuals with disease, who test positive
- false positives (FP) – individuals free of disease, who test positive
- true negatives (TN) – individuals free of disease, who test negative
- false negatives (FN) – individuals with disease, who test negative.

Clearly, the number of individuals with disease equals (TP+FN) and the number without equals (TN+FP). The total number of positive tests is (TP+FP) and of negative (TN+FN).

These data can conveniently be arranged in a matrix (Table 2.6). It is then easy to derive the other important

TABLE 2.6 A matrix for classifying test results

		Test result		Totals
		Positive	Negative	
Disease status	Positive	TP	FN	TP+FN
	Negative	FP	TN	FP+TN
Totals		TP+FP	TN+FN	ALL

TP = true positive; TN = true negative; FP = false positive; FN = false negative.

TABLE 2.7 Important parameters relating to test performance

Parameter	Definition	Formula (expressed as %)
Prevalence	Number of individuals with the disease expressed as a fraction of the population	$\frac{TP+FN}{(TP+FN+TN+FP)}$
Sensitivity	Number of true positives in all individuals with disease	$\frac{TP}{(TP+FN)}$
Specificity	Number of true negatives in all individuals free of the disease	$\frac{TN}{(TN+FP)}$
Positive predictive value	Number of individuals correctly defined as having disease	$\frac{TP}{(TP+FP)}$
Negative predictive value	Number of individuals correctly defined as free from disease	$\frac{TN}{(TN+FN)}$
Efficiency of test	Fraction of individuals correctly assigned as either having, or being free from disease	$\frac{(TN+TF)}{(TP+TF+FP+FN)}$

TP = true positive; TN = true negative; FP = false positive; FN = false negative.

TABLE 2.8 Results matrix using the upper reference limit as cut-off

		Test result		Totals
		Positive	Negative	
Disease status	Positive	TP = 62	FN = 18	TP + FN = 80
	Negative	FP = 11	TN = 109	FP + TN = 120
Totals		TP + FP = 73	TN + FN = 127	ALL = 200

TABLE 2.9 Results matrix using a higher upper reference limit as cut-off

		Test result		Totals
		Positive	Negative	
Disease status	Positive	TP = 50	FN = 30	TP + FN = 80
	Negative	FP = 2	TN = 118	FP + TN = 120
Totals		TP + FP = 52	TN + FN = 148	ALL = 200

parameters relating to test performance, that is, prevalence, sensitivity, specificity, predictive value and efficiency (Table 2.7).

The predictive value model examines the performance of a test in defined circumstances. Clearly, it depends upon being able to determine the numbers of true and false positive and negative results, which depends upon there being an independent, definitive diagnostic test. The importance of the independence of the test under investigation from the definitive test may appear self-evident but, in practice, it is frequently overlooked.

Example

These concepts, and some of the pitfalls inherent in their use, can be illustrated with reference to a hypothetical example. Suppose that it is desired to assess the value of measurements of serum γ -glutamyltransferase activity to assist in the diagnosis of alcohol abuse in patients attending a drug-dependency clinic. It is decided that the test will be regarded as positive if the γ -glutamyltransferase activity exceeds the upper reference limit. Disease status is determined by a validated questionnaire.

Over the course of a year, 200 patients are tested. Enzyme activity is found to exceed the chosen limit in 73, but only 62 are judged to be abusing alcohol on the basis of the rigorous questionnaire. In 18 individuals judged to be abusing alcohol, enzyme activity is not elevated.

The results matrix can thus be completed as in Table 2.8. Calculation then shows that:

$$\text{Prevalence} = (62 + 18) / 200 = 0.400 = 40\%$$

$$\text{Sensitivity} = 62 / (62 + 18) = 0.775 = 78\%$$

$$\text{Specificity} = 109 / (109 + 11) = 0.908 = 91\%$$

$$\text{PV}(+) = 62 / (62 + 11) = 0.849 = 85\%$$

$$\text{PV}(-) = 109 / (109 + 18) = 0.858 = 86\%$$

$$\text{Efficiency} = (62 + 109) / 200 = 0.855 = 86\%$$

Thus, the test correctly identifies 78% of alcohol abusers; further, if a patient has an elevated enzyme activity, there is an 85% probability that he is abusing alcohol. However, it might be considered that a significant number of patients are being misclassified and that better performance might be achieved if the cut-off value for a positive test were to be set at a higher enzyme activity, say twice the upper reference limit.

The results matrix might then appear as in Table 2.9, from which:

$$\text{Prevalence} = (50 + 30) / 200 = 0.400 = 40\%$$

$$\text{Sensitivity} = 50 / (50 + 30) = 0.625 = 63\%$$

$$\text{Specificity} = 118 / (118 + 2) = 0.983 = 98\%$$

$$\text{PV}(+) = 50 / (50 + 2) = 0.961 = 96\%$$

$$\text{PV}(-) = 118 / (118 + 30) = 0.797 = 80\%$$

$$\text{Efficiency} = (50 + 118) / 200 = 0.840 = 84\%$$

The effect has been to decrease the sensitivity of the test, which now only correctly identifies 63% of alcohol abusers. On the other hand, if a patient tests positive, the

probability that he is an alcohol abuser is now 96%. The overall efficiency of the test is little changed.

This example illustrates two important points about the predictive value model. The first is that sensitivity and specificity tend to vary inversely; second, appropriate selection of the criterion for positivity or negativity (here the level of enzyme activity) allows either one to be maximized. If the cut-off is set very high, positive results will only occur in individuals with disease; there will be no false positives and specificity will be 100%, although sensitivity will be low. On the other hand, if the cut-off is set low, no cases will be missed, but the false positive rate will be high; sensitivity will be 100%, but specificity will be low. This is illustrated in Figure 2.3.

Whether it is desirable to maximize sensitivity, specificity or efficiency depends on the nature of the condition being investigated, as discussed below.

Prevalence and predictive value

While sensitivity and specificity are dependent on the characteristics of the test and the nature of the condition being investigated, predictive values are dependent on the prevalence of the condition in the population under examination. Consider the application of the measurement of γ -glutamyltransferase activity to identify alcohol abuse during pregnancy. If 500 women were screened and the cut-off taken as the upper limit of the reference range, then the results matrix might show as in Table 2.10, from which:

$$\begin{aligned} \text{Prevalence} &= (24+7)/500 = 0.062 = 6\% \\ \text{Sensitivity} &= 24/(24+7) = 0.774 = 77\% \\ \text{Specificity} &= 426/(426+43) = 0.908 = 91\% \\ \text{PV}(+) &= 24/(24+43) = 0.358 = 36\% \\ \text{PV}(-) &= 426/(426+7) = 0.983 = 98\% \\ \text{Efficiency} &= (426+24)/500 = 0.900 = 90\% \end{aligned}$$

The sensitivity and specificity are unchanged compared with the previous example, but, because of the much lower prevalence of alcohol abuse in this group, the predictive value of a positive test is low: only 36% of those with a positive test are alcohol abusers. At the same time, the predictive value of a negative test is higher than in the patients attending the drug dependency clinic: this is again a consequence of the lower prevalence in pregnant women. It is instructive to note that the efficiency of the test appears greater in this context (90% in comparison to 84% in the drug abuse patients). Overall, the proportion of tests that correctly assign patients to the alcohol abuse or non-abuse categories is higher, although its performance in diagnosing abuse alone is poorer.

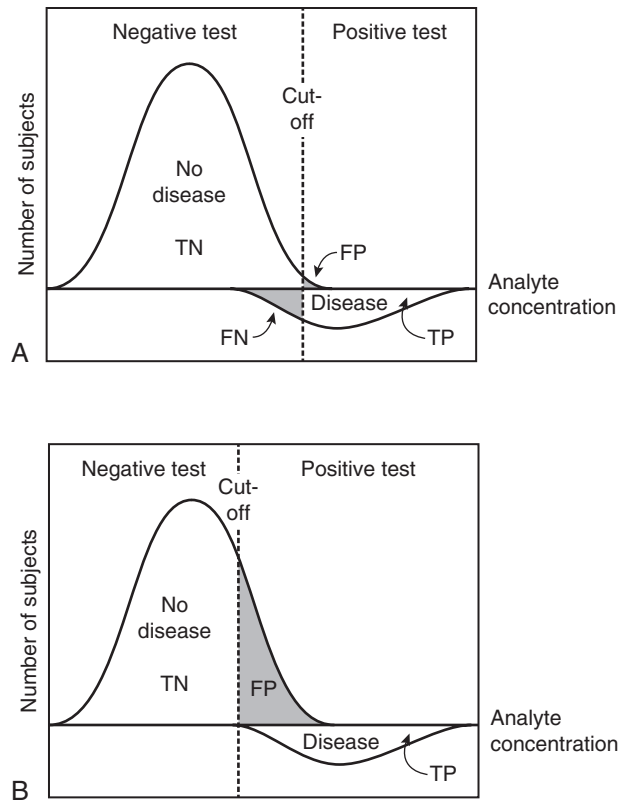


FIGURE 2.3 ■ The effect of moving the cut-off point that determines positivity/negativity of a test result. Hypothetical distributions for the concentrations of an analyte with and without disease are shown. Because these overlap, if the cut-off is selected to decrease the number of false positive (FP) results (and hence increase specificity), (A) there are a significant number of false negative (FN) results (decreasing the sensitivity). If the cut-off is set lower (B), false negatives are eliminated (maximizing sensitivity), but at the expense of increasing the number of false positives (and decreasing specificity). The distribution for individuals with disease has been shown below the axis for clarity. TP, true positive; TN, true negative.

The effect of disease prevalence on predictive values is an important one. It means that the performance of a test in one group of patients cannot be translated to a different group. Although the predictive value of a new test may appear high in the typical experimental setting where equal numbers with and without disease are tested (the prevalence is 50%), it is likely to be much lower 'in the field'. Thus, the performance of a new test must be analysed in groups of individuals comparable with those in whom it is to be used in practice. In the past, many exaggerated claims for tests have been made on the basis of their performance in a highly selected group.

TABLE 2.10 Results matrix in a population with a lower disease prevalence

		Test result		Totals
		Positive	Negative	
Disease status	Positive	TP=24	FN=7	TP+FN=31
	Negative	FP=43	TN=426	FP+TN=469
Totals		TP+FP=67	TN+FN=433	ALL=500

Practical applications of the predictive value model

High sensitivity in a diagnostic test is required when a test is being used to diagnose a serious and treatable condition. The false negative rate (missed diagnoses) must be as low as possible, even though this inevitably means that false positives will occur. Individuals testing positive will be subjected to further, definitive tests and, provided that these cause no harm to those who test positive and turn out not to have the condition, this may be an acceptable price to pay for making the diagnosis in all those who have the condition. The programmes for neonatal screening for phenylketonuria and other harmful conditions in the newborn exemplify tests requiring maximum sensitivity. In this context, it is instructive to note that although the sensitivity of the screening test for phenylketonuria is 100%, and its specificity approaches this value, because the condition has such a low prevalence (approximately 1 in 10 000), the predictive value of a positive test is only of the order of 10%.

High specificity (no false positives) is desirable in a test to diagnose a condition that is severe, but either it is untreatable or the benefits of treatment are unpredictable, when knowledge of its absence is potentially beneficial. Multiple sclerosis is frequently cited as the classic example of such a condition. High specificity is also required when it is desired to select patients with a condition for a trial of some new form of treatment. If individuals without the condition were to be included in the treatment group, the results of the trial (assuming that the treatment is in some measure effective and does no harm to those without the disease) would give a false view of its efficacy.

Diagnoses are rarely made on the basis of single tests. Ideally, a single test or combination of a small number of tests should be used to identify individuals in whom the probability of a condition is significantly higher than in the general population (the prevalence is higher), who can then be further investigated. The initial tests can be simple, cheap and unlikely to do harm, but the further tests may be more elaborate, expensive and possibly associated with some risk. In essence, if the performance of individual tests is known and the objectives of testing can be precisely defined, the predictive value model can be used to determine the appropriate sequence or combination of tests to achieve the desired objective.

Receiver operating characteristic curves

Another use of the predictive value model is to compare the performance of two tests. This can be done by determining sensitivity and specificity for each test using a range of cut-off limits to define positivity and plotting one rate against the other. The resulting curves are known as receiver operating characteristic (ROC) curves. The reader is warned that four variants of ROC curves may be encountered, according to the way the data are projected on the axes, but this does not affect their interpretation. Figure 2.4 shows hypothetical ROCs for two tests, A and B, each applied to the same group of individuals to make the same diagnosis.

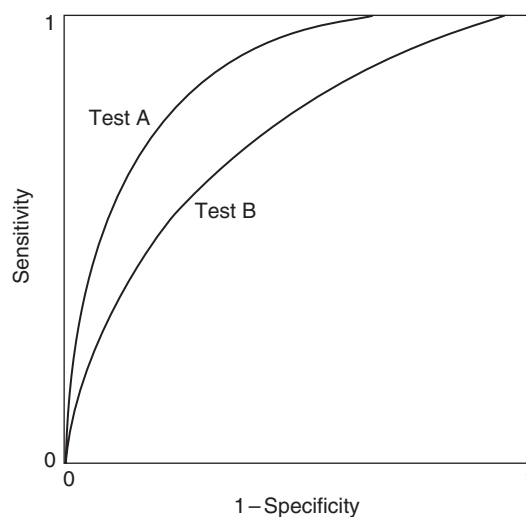


FIGURE 2.4 ■ Receiver operating characteristic curves for two tests being assessed for the diagnosis of the same condition in the same patients. Test A gives better specificity at any level of sensitivity.

It is clear that test A gives the better discrimination, since for any level of sensitivity its specificity is superior. This may be useful information, but these are by no means the only factors that may need to be taken into account in choosing a test. The economics and practicalities of the tests will also be of relevance to which one is chosen.

The interpretation of the ROC curves in Figure 2.4 is straightforward. In practice, the curves for two tests may cross; the relative performance must then include assessment of the areas under the curves. The greater the area, the greater the efficiency.

Likelihood ratios

Although the predictive value model can be used to determine the most appropriate cut-off for a test result for clinical purposes, it can only classify results as positive or negative, and does not give any indication of the degree of abnormality, which may be of prognostic significance.

In contrast, likelihood ratios can be applied to individual test results. The likelihood ratio states how much more likely a particular test result (or combination of results) is to occur in an individual with a disease than without it (positive likelihood ratio, LR(+)), or not to occur, again in individuals with a disease than without it (negative likelihood ratio LR(-)). The ratios are calculated thus:

$$\text{LR}(+) = \text{sensitivity} / (1 - \text{specificity})$$

$$\text{LR}(-) = (1 - \text{sensitivity}) / \text{specificity}$$

To give an example, an LR(+) of 10 implies that the given result will be obtained ten times more frequently in a population with the condition in question than in one without. A LR(-) of 0.01 implies that the given result will be obtained 100 times more often in individuals without the condition than in those who have it. Calculation of the LR(+) and LR(-) allows the post-test probabilities for

either a positive or negative result to be calculated. The post-test probability can be calculated as follows:

$$\text{Post-test probability} = \frac{(\text{pre-test probability} \times \text{LR})}{(1 + \text{pre-test probability} \times (\text{LR} - 1))}$$

The pre-test probability is equal to the prevalence of the condition in comparable individuals (e.g. a population being screened).

Positive and negative likelihood ratios can be combined to give the diagnostic odds ratio (DOR) for a particular test result:

$$\text{DOR} = \text{LR}(+) / \text{LR}(-) = (\text{TP} \times \text{TN}) / (\text{FP} \times \text{FN})$$

The DOR expresses the odds of a result being positive in a patient with the condition being investigated compared with one who does not have the condition, and is a measure of the overall accuracy of a test.

CONCLUSION

It is a simple matter to fill out a request form for a biochemical test, but before doing so, it is essential to consider why the test is being done and for what the results will be used. Many factors can affect the results of laboratory tests, apart from the pathological process that is being investigated. Some of these are obvious, others less so. All need to be minimized by careful patient preparation, sample collection and handling, analytical procedures and data processing, if the results are to be reliable and fit for whatever purpose they are to be used.

The interpretation of biochemical data requires adequate knowledge of all the factors that can affect the test result. These include the physiological, biochemical and pathological principles on which the test is based, as well as the reliability of the analysis. It is also necessary to understand the statistical principles that relate to the distribution of data in healthy individuals and those with disease, and how the test performs in the specific circumstances in which it is used. Such knowledge is

also essential for the appropriate selection of laboratory investigations.

ACKNOWLEDGEMENT

We would like to thank William Marshall, who wrote this chapter for previous editions of the book.

Further reading

Cerretti F, Hinzmann R, Panteghini M. Reference intervals: the way forward. *Ann Clin Biochem* 2009;46:8–17.

A good review of current and potential methods to establish reference intervals.

Fraser CG. *Biological variation: from principles to practice*. Washington: AACC Press; 2001.

An updated exposition of some of the topics discussed in the same author's book: Fraser CG. Interpretation of clinical chemistry laboratory data. Oxford: Blackwell Scientific; 1986. These include a detailed discussion of biological variation and its consequences for analytical performance, reference ranges and analytical goals. However, the earlier book is likely to be more accessible – albeit less comprehensive – to readers who are uneasy with statistics and mathematics.

Fraser CG, Fogarty Y. Interpreting laboratory results. *Br Med J* 1989;298:1659–60.

Although written some 23 years ago for a general medical readership, this leading article is still worth reading. It was one of the first general articles to draw attention to the importance of taking analytical and biological variation into account when interpreting laboratory data.

Galen RS, Gambino SR. *Beyond normality: the predictive value and efficiency of medical diagnoses*. New York: John Wiley; 1975.

The seminal text on the predictive value concept.

Jones R, Payne B. *Clinical investigation and statistics in laboratory medicine*. London: ACB Venture Publications; 1997.

An accessible text, which manages successfully to explain statistical topics without overwhelming the reader with symbols and mathematics.

Ricos C, Alvarez V, Cava F. Essay: biologic variation and desirable specifications for QC, <http://www.westgard.com/guest17.htm>.

A comprehensive list of data on biological variation and related matters, based upon the original article.

Ricos C, Alvarez V, Cava F et al. Current databases on biological variation: pros, cons and progress. *Scand J Clin Lab Invest* 1999;59:491–500.

Swinscow TDV. *Statistics at square one*. 9th ed 1997. Available for free online at: <http://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one>.

An excellent statistics book that begins from first principles.

United Kingdom National External Quality Assessment Service (UK National External Quality Assurance Scheme), <http://www.ukneqas.org.uk/content/PageServer.asp?S=995262012&C=1252&CID=1&type=G>.

Quality aspects of laboratory medicine

Helen Bruce • Marta Lapsley

CHAPTER OUTLINE

INTRODUCTION 21

WHAT IS QUALITY? 21

QUALITY STANDARDS 21

Quality assurance 21

Regulation of laboratories 22

Quality management systems 22

Personnel 22

Premises and environment 22

Information systems 23

Evaluation and audit 23

CLINICAL QUALITY INDICATORS 24

Clinical effectiveness 24

Key performance indicators 24

Demand management 25

EVIDENCE-BASED CLINICAL BIOCHEMISTRY 25

POINT-OF-CARE TESTING 25

CONCLUSION 26

INTRODUCTION

It was emphasized in Chapter 2 that every step in the process, from the ordering of the investigation, through collection of the specimen(s) required, their transport to the laboratory and analysis, to the delivery of a report to the clinician, is vital in ensuring that results of investigations are used appropriately. Interwoven throughout each step is the requirement to ensure the quality of the whole process. Chapter 2 has already discussed the need to minimize analytical variation through the use of processes to reduce imprecision and bias. The use of standard operating procedures has also been alluded to, but this chapter will expand on the use of laboratory standards to maintain quality.

WHAT IS QUALITY?

Quality is difficult to define, and may be considered to be 'doing the right test, for the right patient, at the right time for the right reason', but may equally be defined as 'adhering to stringent processes in order to complete the task'. Ultimately, perhaps the question to ask is how to assess the quality of a laboratory service.

QUALITY STANDARDS

Quality assurance

Clinical biochemistry laboratories operate extensive quality assurance procedures to ensure that the results

that they produce can be relied upon to support clinical decision-making. They include internal quality control (IQC) schemes, for example involving replicate analysis of clinical samples and the repeated analysis of reference samples of identical composition, and external quality assurance (EQA) schemes involving the analysis of identical samples by a large number of laboratories.

The measurement of IQC samples allows a prospective estimate of the precision of the analyses that are being performed. The frequency of measurement of such controls may depend upon each laboratory's workload and the type of analysis undertaken. For example, it is not unusual to run quality control material at the start and the end of a sample batch measured by enzyme-linked immunosorbent assay (ELISA).

The measurement of EQA samples provides laboratories with a retrospective view of the accuracy of their measurements and the consistency of any bias that may be present. The data can be invaluable when troubleshooting possible assay problems, by allowing the laboratory to compare their result with both the reference value (if available) and the method-specific mean. The review and interpretation of EQA reports is an important skill for laboratory personnel to develop because of the variety of ways in which different EQA schemes choose to present their data. (Examples may be accessed via the website addresses of the various EQA providers included in Further reading, below.)

Not only the analysis itself, but also the interpretation of the result should be subjected to quality assurance: there exists a peer-reviewed EQA scheme for interpretative comments.

Reviewing quality assurance procedures is an essential part of the audit of laboratory performance (the assessment of overall performance with regard to the appropriateness of the use of tests, the interpretation of their results, adherence to standard procedures, cost-effectiveness etc.). Laboratory data should also be included in the medical audit, in which the effectiveness of all aspects of clinical care is examined. The ultimate test of the performance of laboratory data is provided by the clinical outcome when this has been determined, in whole or in part, by those data.

Regulation of laboratories

External regulation of laboratories has evolved over some time to include every aspect of the laboratory service. Independent regulatory bodies are empowered to award accredited status to laboratories meeting stringent standards and to fail to award or remove accreditation from laboratories which are found to have serious failings at inspection. Throughout the world, there are several organizations that fulfil this role: International Standards Organization (ISO), United Kingdom Accreditation Service (UKAS) and the Division of Laboratory Services and Standards in the USA, to name but a few. Their main aim is to ensure a laboratory's quality and competence to perform specified tasks. It is important to note that all parts of the laboratory are inspected, including the environment in which the laboratory is based. Very dilapidated buildings (in which pathology departments are sometimes located) will be a cause of concern for the inspectors, as it may reflect the low priority given to laboratories within a wider healthcare setting, such as in a hospital.

Within the UK, for example, the current list of standards against which laboratories are judged comprises almost 60 pages. The broad areas covered are listed in [Box 3.1](#). It is worth noting that the analytical process itself is just one out of eight categories examined by such an inspection.

Quality management systems

A system of 'Quality Management' is an all-encompassing term used to describe the steps a laboratory may take to maintain current procedures and to establish continual service improvement via quality objectives.

The majority of laboratories in the UK will have a designated Quality Manager, either exclusively within

their own discipline or as part of a multidisciplinary team. The development of web-based document management systems (e.g. QPulse, i-Passport) has enabled laboratories to greatly improve document control and to track review dates, changes and the number of printed copies in circulation. Complete written procedures should be in place for all tasks performed in a laboratory. Manufacturers are obliged to publish 'Instructions for Use' with each reagent supplied that contains the majority of worthwhile data. These procedures, however, form only part of the process, as the essence of a quality management system is that it documents what should take place within the laboratory in its entirety. All members of laboratory staff should be familiar with the quality management system, understand the requirement for such a system and the need to maintain the documents within it. Indeed, it is the effective implementation of the system, and adherence to it, that is likely to lead to quality improvement. Monitoring the effectiveness of the system is achieved by audit (see below) and by reviewing processes in a regular basis.

Many laboratories try to minimize the amount of printed material generated, owing to ever greater space and storage constraints. Improved reliability of information technology and the need to minimize waste are also important driving factors towards operating a paperless Quality Management system.

Personnel

When contemplating quality, staffing issues may not immediately spring to mind. Staff are, however, the most valuable assets of any laboratory. Staff morale and continuing professional development play a vital role in a 'good' quality laboratory. Education and training of all levels of staff within the laboratory ensures that their knowledge and understanding of laboratory science is current and up-to-date. The competence of staff undertaking laboratory tasks should be documented and regularly reviewed. Formal individual review meetings are recommended to assess training needs and to promote further development.

Premises and environment

The control of access to laboratories and the efficient use of heating, lighting and space are important considerations when evaluating a suitable working environment. Health and safety aspects of the premises and environment need to be reviewed, both in terms of staff safety (e.g. lone working, risk of slips and trips, adequate fire plan) and general laboratory safety such as electrical issues (e.g. overloaded sockets, regular testing of electrical appliances, use of extension cables) and chemical safety (e.g. appropriate use of material safety data sheets and adequate storage facilities for acids, controlled drugs and toxins). The latter category may also include awareness of laboratory waste disposal, which will encompass both biological and chemical waste. In the UK, for example, NHS Trusts have financial incentives related to recycling and alternative methods of disposal other than incineration.

BOX 3.1

The areas examined by the UK Accreditation Service

- Quality management system
- Personnel
- Premises and environment
- Equipment, IT and materials
- Pre-examination processes
- Examination processes
- Post-examination phase
- Evaluation and quality assurance

Information systems

The development of sophisticated information systems has had a huge impact on numerous aspects of laboratory medicine. Pre-analytical improvements include the ability to display relevant test information and sample requirements via electronic requesting. Real-time availability of results and the possibility of report acknowledgement are also areas revolutionized by information systems.

Information systems affect the way in which laboratories communicate with their users. Many laboratories now maintain an electronic laboratory handbook, as well as a paper copy. The advantages of an electronic version are that it reduces the cost of production (and paper waste) and it is easier to maintain and keep up-to-date. Patient information leaflets and testing protocols such as dynamic function tests can easily be accessed via hyperlinks, if available.

The increasing use of e-mail has allowed laboratories to communicate electronically with end-users, although security must be assured before sending patient identifiable data. Future developments will increasingly allow patients to use this or other electronic means to access their own results directly. In the UK, some patient groups, such as those with kidney disorders, already have electronic access to selected results via a nationally organized database.

The integration of laboratory information systems with electronic patient records in both primary and secondary care has revolutionized the delivery of pathology results over the past decades. The ability to view test results remotely and securely has improved the safe delivery of healthcare, especially in non-traditional settings, and has expedited the move of care from hospitals to within the community.

Evaluation and audit

The use of audit tools in laboratory medicine is widespread. During formal assessment of the laboratory's service, a combination of vertical, horizontal and examination audits are used to assess compliance with the quality standards set. Vertical audits generally look at the sample journey throughout the entire testing process, from test requesting and phlebotomy (if possible), to report generation and receipt by the requestor. Actual events are compared with written procedures to check adherence. A failure to follow any stated procedure may trigger a horizontal audit of that particular part of the testing process. A rather extreme example would be that during the course of a vertical audit, it is discovered that no quality control material was run during analysis of a particular analyte such as glycated haemoglobin; examiners may act upon this information and perform a horizontal audit of all recent analytical batches of glycated haemoglobin to see if this was an isolated occurrence or if there was a systematic failure to follow the stated procedure.

Ensuring that the audit cycle is complete is critical for continued quality improvement. Re-auditing previous processes, in which highlighted deficiencies have forced a change in practice, is vital to ensure that the changes in practice have resulted in the anticipated improvements. [Figure 3.1](#) shows the key steps to completing an audit cycle.

Clinical audit is another key aspect of evaluating laboratory effectiveness. Clinical audit plays an important role

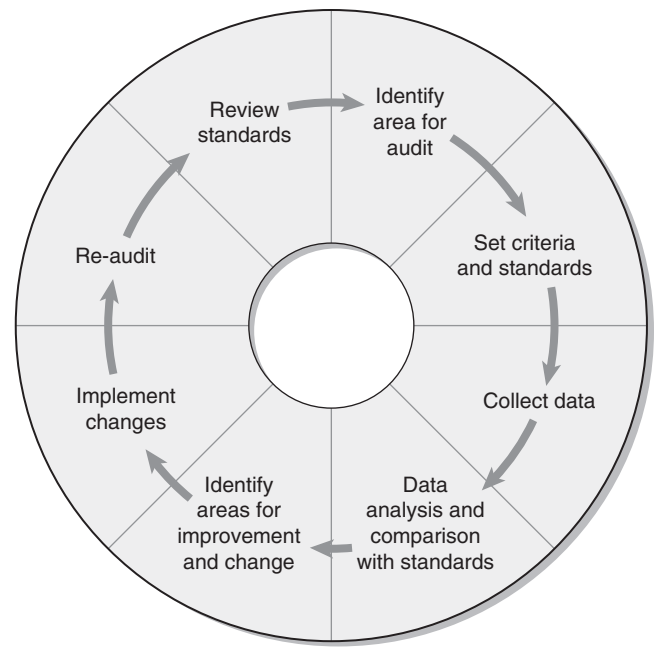


FIGURE 3.1 ■ The complete audit cycle. The process can be started at almost any point of the cycle, but often the first task is to identify an area of interest and then determine standards against which to conduct an audit.

in trying to assess the impact of laboratory testing on patient pathways and outcomes. Ultimately, the role of the laboratory is to improve the healthcare of patients and, although difficult to measure, the contribution of laboratory testing should be documented. Collaboration with other clinical disciplines allows the laboratory to forge vital relationships with clinicians and other staff within the healthcare environment.

The third type of evaluation that the laboratory should undertake is an assessment of user satisfaction. It is important to engage all users of the laboratory service, as each different group may have different needs and requirements. Clinicians, nurses, other healthcare professionals and other laboratories should be asked to contribute. All aspects of the service can be covered from the users' opinions on the judicious use of interpretative comments to turnaround times and phlebotomy services. User satisfaction surveys may eventually be circulated to patients (who may be the true end-users of the service), especially those with long-term chronic disease states (e.g. diabetes mellitus, chronic kidney disease), as they take greater ownership of their health and treatment targets. [Box 3.2](#) gives a list of the other areas covered by questions commonly seen in user surveys.

BOX 3.2

Examples of topics covered within user satisfaction surveys

- Availability of clinical advice
- Ease of requesting
- Availability of test information
- Frequency of phlebotomy collections (primary care)
- Online access to results (availability of)
- Out of hours service provision

CLINICAL QUALITY INDICATORS

Laboratories are generally adept at maintaining quality standards within their own environment; however, the need to determine and measure objectively their clinical effectiveness is growing, owing to increasing pressure on limited resources. Clinical quality indicators have recently been developed to aid in this process. The key to the use of these indicators is that they must be easily measurable and that the results can be collected in a timely manner (a yearly review would be of limited use in highlighting errors and improving performance promptly). Several quality indicators can be brought to mind, but if the mechanism for capturing the necessary data is too demanding or difficult to achieve, then the whole process becomes meaningless.

Quality indicators can fall into two broad categories: those that are designed to highlight improvement and those that are used to capture a possible deterioration in standards. An example of each may include the use of external quality assurance to show improving assay precision and the monitoring of turnaround times from emergency departments to detect peaks and troughs in service provision. Turnaround times in particular, can be an area where the laboratory and requesting clinicians can have differing views, with clinicians considering turnaround time to be 'vein-to-brain' time rather than from receipt of sample in the laboratory to a clinically validated result. [Table 3.1](#) lists other potential quality indicators that laboratories may be able to use to monitor their own performance.

Clinical effectiveness

In the context of laboratory medicine, this may be perceived as the laboratory's ability to introduce new tests and to dissuade clinicians from using tests deemed to be obsolete. Clinical biochemists may be involved in the development of specific testing protocols in Emergency Departments, for example, or in the interpretation of data in complex cases. Another way in which laboratories contribute to effective clinical care, is the identification of artefactual causes of altered biochemical analytes, such as the effect of thrombocytosis on potassium concentration. The implementation of national guidelines and the adoption of agreed testing strategies or more precise analytical methods (such as the measurement of creatinine by enzymatic methods), are all examples of ways in which the clinical effectiveness of laboratories has been improved.

Key performance indicators

In an attempt to objectively assess the quality and effectiveness of predominantly pre- and postanalytical processes within the laboratory, the Royal College of Pathologists in the UK has established Key Performance Indicators. These suggested standards cover a wide range of areas, from the percentage of staff in training to the speed at which a laboratory can respond to a request for clinical advice (what defines clinical advice may be a contentious issue). [Table 3.2](#) lists some of the other additional performance standards suggested by the Royal College of Pathologists. These are still in their infancy and some

TABLE 3.1 Examples of potential quality indicators

Category	Example quality indicator
Preanalytical	Investigation strategies and protocols
Analytical	Introduction and maintenance of reference (critical) change values
Postanalytical	Matching analytical performance to clinical requirements
	Data interpretation
	Addition of appropriate further tests
Direct clinical services	Patient/ward interventions, e.g. review of patients with metabolic disease, electrolyte or fluid balance problems
	Participation in multidisciplinary team meetings
Education	Undergraduate and postgraduate teaching; participation at hospital-wide events such as Grand Round meetings; newsletters

TABLE 3.2 Some key performance indicators recommended by the UK Royal College of Pathologists

Category	Performance standard and method of measurement
Training	15–30% of the biomedical scientist, clinical scientist and medical staff should be in training posts
User surveys	The laboratory should include patients in the annual user satisfaction survey
Turnaround times	The percentage of investigations from A&E to be completed within 1 h. Target: 90% from April 2014
Results communication	The percentage of critical results phoned to the requesting clinician within 2 h of availability. Target: 97% by April 2014

will require further clarification. However, they do provide an excellent basis on which a laboratory can assess its current performance. In order for these key performance indicators to have an impact on commissioning patterns and thereby possibly reward the best-performing laboratories, each participating laboratory should be expected to publish some of their results on a regular basis, thus allowing comparison with their peers.

Demand management

In the past, clinicians were often allowed unrestricted access to all available assays within the laboratory, but as costs increase, test repertoires expand and undergraduate training in laboratory medicine decreases, laboratories now find themselves attempting to manage the increasing demand placed on them. Considerable work has been done to ascertain sensible repeat testing intervals for analytes which are too frequently requested in secondary care. Some examples include bone and liver profiles, vitamin D and thyroid function testing. Electronic order requesting enables the laboratory to communicate warning messages to the clinician prior to further repeat requesting and also to educate the users about the most efficient use of laboratory services. Another form of demand management may involve the manual vetting of requests from within the laboratory. This may be for specialist tests that are referred to other laboratories or for analytes that have a specific clinical role, such as tumour markers. Vetting performs a dual purpose; first, it enables the lab to control costs and second, it allows the lab to open a dialogue with the requestor regarding the need and suitability of the test.

EVIDENCE-BASED CLINICAL BIOCHEMISTRY

The later years of the 20th century saw the establishment and development of the concept of evidence-based medicine, in essence the use of the best available data from robust clinical studies to inform the management of patients, including diagnosis, prognosis and treatment. This has led to the promulgation of the concept of evidence-based clinical biochemistry (and evidence-based laboratory medicine in general) in which, for example, the relationship between investigations and clinical outcomes is rigorously analysed, using tools such as likelihood ratios (see p. 19), to determine which investigations are the most appropriate (in terms of being valid and informative) in individual clinical situations and evaluating their subsequent use.

In evidence-based medicine, the gold standard for providing evidence is the systematic review of evidence adduced from clinical trials, preferably randomized and double blinded. Such reviews should be designed to minimize bias, random errors and confounding. Systematic reviews may include meta-analysis – a statistical analysis of the results from a series of independent studies designed to produce a single statement about outcome, whether it be the efficacy of a particular treatment or the value of a particular diagnostic investigation.

Although the number of systematic reviews of diagnostic investigations published annually is increasing,

there are none for many well-established biochemical tests, and there is no requirement to conduct studies of the efficacy and validity of new tests before they are introduced to the market, as there is with therapeutic agents. For example, numerous new tumour markers have been developed in recent years, but for many, there is only limited evidence (and only in specific contexts) of their clinical value. The context in which an investigation is used is an essential consideration. For example, there is no doubt that the measurement of prostate-specific antigen (PSA) is of value in monitoring the response of patients with prostatic cancer to treatment, but basal values have little prognostic significance and there remains no definite evidence that screening older men for prostatic cancer by measuring PSA has any effect on outcome.

Evidence-based clinical biochemistry is gathering momentum. It starts with the technical evaluation of a test and encompasses the assessment of diagnostic performance and clinical utility, as discussed in Chapter 2. One of its important products is the development of guidelines for the investigation of patients with specific conditions. But just as there is continued clinical surveillance of drugs after they have been introduced to the market (resulting, in some cases, in their later being withdrawn or having their licences altered), so the impact of guidelines and other products of evidence-based laboratory medicine must be reviewed to ensure that they are appropriate and effective, and remain so with the passage of time. The process of clinical audit – defining standards; examining processes against these standards; identifying deficiencies; modifying the processes to correct them, and repeating the cycle to check for improved performance – is fundamental to this.

POINT-OF-CARE TESTING

There is a current trend to deliver a more patient-centric service with some laboratory testing taking place within the community, at point-of-care. This desirability, under some circumstances, to provide biochemical services at the bedside or in the clinic or doctor's surgery, has led to the development of techniques and analytical instruments that can be used reliably by individuals who do not necessarily have rigorous laboratory training. Examples include hydrogen ion and blood gas analysis in intensive therapy units, principally as an aid to the management of patients on ventilators; dipstick analysis of urine samples; blood glucose measurements on hospital wards and in patients' homes, and bilirubin analysis in neonatal intensive therapy units. The advantages of such testing include the rapid availability of results at the location where management decisions are made and an ability to perform the test frequently; both these facilitate rapid responses to changes in the patient's condition. Certain sources of error may be eliminated (although care must be taken not to introduce others) and there may be financial savings, for example in relation to the transport of specimens or of patients; however, the cost per test is frequently much higher for point-of-care tests (PoCTs) than for the central laboratory.

This is not a new concept. In Victorian times, physicians would perform simple pathological tests at the bedside and it was the increasing complexity of the tests that led to the setting up of centralized laboratory facilities. Point-of-care testing offers considerable advantages over laboratory-based testing under some circumstances, as is apparent from the examples cited above. However, it is important that the tests and instruments are reliable; that only staff who have been trained to use them are allowed to do so; that the results are subject to appraisal through a quality assurance scheme (preferably supervised by a clinical biochemist) and that there is adequate skilled technical and analytical back-up, should problems arise. Laboratories are expected to be at the forefront of these developments and work with the PoCT user to establish a set of quality standard analogous to those used in the main laboratory. The international standard, ISO 22870, sets out the specific standards that should be met when such testing takes place within a hospital environment.

There are however, ongoing clinical governance issues regarding the use of PoCT. Two key areas of concern are the assimilation of the PoCT result into the patient record and the need to ensure adherence to rigorous testing processes (a complete audit trail). As an example, the latest blood glucose meters available on the market have the capability of connecting to both laboratory and hospital information management systems. These meters control access by requiring unique identification of individual users. Additionally, connected meters can be controlled remotely from the laboratory to ensure that the meter has had quality control checks prior to measurement of patient samples. Remote access to these meters can also ensure that use by untrained individuals is prevented.

These developments are excellent examples of the progress that has been made by manufacturers in this area. The need to establish good working relationships with the manufacturer and colleagues that use the system is vital to ensure success.

As the use of PoCT devices becomes more prevalent, regulation will become stricter and it is envisaged that the laboratory will take a greater role in their provision and

management, possibly through service level agreements with community practitioners.

CONCLUSION

Laboratories are obliged to provide a good-quality service, which should be independently assessed and regulated. As quality standards have evolved, so has the need to include inspection of procedures to cover the entire sample journey, from phlebotomy to receipt of the final report. Laboratories should review the needs of their end-user, be it clinician, nurse or patient and try to adapt their service accordingly.

The need to monitor the clinical effectiveness of laboratories to demonstrate continuous quality improvement and better patient outcomes has resulted in the development of clinical quality indicators. These indicators should be used to provide laboratories with sufficient data to identify areas of good practice and areas where further improvement is necessary and to promote a culture where continuous quality improvement is integrated into routine activity.

FURTHER READING

Barth J. Clinical quality indicators in laboratory medicine. *Ann Clin Biochem* 2012;49:9–16.

Barth J. Selecting clinical quality indicators for laboratory medicine. *Ann Clin Biochem* 2012;49:257–61.

Moore AR. Evidence-based clinical biochemistry. *Ann Clin Biochem* 1997;34:3–7.

A concise introduction to this topic.

National External Quality Assurance Scheme, UK NEQAS. www.ukneqas.org.uk.

Oxford Centre for Evidence-based Medicine. <http://www.cebm.net/?0=1162>.

The website of the Oxford Centre for Evidence-based Medicine. This page illustrates the use of likelihood ratios to modify diagnostic probabilities, but there are many other useful pages.

Price CP, Christensen RH, editors. Evidence-based laboratory medicine: principles, practice and outcomes. 2nd ed. Washington: AACC Press; 2007.

A series of essays discussing the acquisition and use of evidence for diagnostic investigations, clinical audit and the development and evaluation of guidelines.

Randox External Quality Assurance Scheme, RIQAS. www.riqas.com.

Royal College of Pathologists. How to assess the quality of a pathology service. London: Royal College of Pathologists; 2011.

United Kingdom Accreditation Service. www.ukas.com

Welsh External Quality Assurance Scheme, WEQAS. www.weqas.com

Sodium, water and potassium

Michael D. Penney

CHAPTER OUTLINE

PHYSIOLOGY 27

- Introduction 27
- Extracellular fluid and sodium 28
- Intracellular fluid and water 30
- Extracellular fluid, intracellular fluid and potassium 32

DISORDERS OF SODIUM METABOLISM 33

- Sodium deficiency 33
- Sodium excess 36

DISORDERS OF WATER METABOLISM 39

- Polyuria 39

- Nocturnal polyuria 43

- Hypernatraemia 44
- Hyponatraemia 46

DISORDERS OF POTASSIUM METABOLISM 52

- Hypokalaemia 52
- Hyperkalaemia 57

CONCLUSION 61

APPENDICES 62

PHYSIOLOGY

Introduction

Water is the most abundant molecule in the human body and is the major solvent (the only other of consequence is fat). The physiological control of the composition and distribution of the fluid spaces is a highly sensitive and complex homeostatic process necessary to maintain a constant milieu interieur. Two main fluid spaces exist: the intracellular fluid (ICF) and the extracellular fluid (ECF). The latter is further separated into the intravascular space (plasma), the interstitial space (which includes lymph) and transcellular fluid (pleural, pericardial, peritoneal, cerebrospinal and gastrointestinal fluids), formed by the transport activity of cells. [Table 4.1](#) summarizes the water content of the body and the distribution of fluid between the main body spaces; the proportion of body water to total body weight is affected by age, gender and body fat content.

The electrolyte compositions of the ECF and the ICF are different – in essence, the extracellular space is a predominantly sodium-containing solution and the intracellular space a potassium-containing solution ([Table 4.2](#)). This fundamental difference in electrolyte composition is maintained by cell membrane transport pumps (energy-consuming ATPases). Electrolyte and protein concentrations of blood are now most commonly measured in serum. In this chapter, unless specified, the term *plasma* is used for describing in vivo concentrations and the term *serum* for measured in vitro concentrations.

Body water moves between the main body spaces through water channels (aquaporins), predominantly under the influence of the osmotic pressures resulting from

dissolved particles in the ECF and ICF on either side of the cell membrane. Under steady-state physiological conditions, the osmotic pressure of the ICF equates exactly with the plasma osmotic pressure. Osmolality represents the molal concentration of solute in a litre of solvent (water) and is expressed as mmol/kg, as opposed to a molar (or calculated osmolar) solution, which is the concentration in the space of a litre of solution (which includes solute space) and is expressed as mmol/L. This apparent nicety in definition does have its uses in differentiating certain real from apparent electrolyte disorders (see [p. 46](#)).

The measured osmolality in ECF cannot, however, always be equated with transcellular osmotic pressure. The cell membrane is selectively permeable to a variety of solutes, but certain natural solutes such as urea, or exogenous solutes such as alcohol, are freely permeable. Thus, an increase in plasma osmolality due to sodium implies an increase in osmotic pressure across the cell membrane, tending to withdraw water from the cell to equalize osmolalities. However, an increase in plasma osmolality due to urea does not have this effect because of the free permeability of urea between the ICF and ECF. This leads to the concept of *effective osmolality or tonicity*, which, under physiological conditions, is primarily dependent on plasma sodium concentration, but under pathological or iatrogenic conditions, may be dependent on other solutes, for example, the effect of glucose in untreated diabetes mellitus or mannitol following its intravenous infusion (used therapeutically precisely for this effect).

The distribution of water between different body spaces is thus dependent on the permeability of the relevant membrane barriers to water, and the quantity of

TABLE 4.1 Body fluid distribution in relation to age and gender

	Infant age 1 year	Adult male age 40 years	Adult female age 40 years
Weight (kg)	7	70	60
Total body water (L)	4.90	42	30
ICF volume (L)	3.15	28	18
ECF volume (L)	1.75	14	12
of which intravascular (L)	0.35	2.8	2.4

TABLE 4.2 Representative molal concentration of electrolytes within the body fluid spaces

Other predominant intracellular counter anions are sulphates and proteinates

Electrolyte	ECF (mmol/kg)	ICF (mmol/kg)
Sodium	152	10
Potassium	4.3	160
Calcium	2.7	1.0
Magnesium	1.1	13
Chloride	109	10
Bicarbonate	29	10
Phosphate	1.5	50

solute within each space. Because water is freely permeable across all cell membranes (except some highly specialized membranes in the nephrons and sweat glands), the water content of body spaces is dependent on the solute content of the space. The Gibbs–Donnan effect is an important force that influences the distribution of solutes. If the barrier separating two compartments is permeable to water and small ions, but impermeable to large ionized molecules, and the larger molecules are confined to one compartment, the concentration of small ions will differ between compartments at equilibrium, and the compartment containing the protein will exert an osmotic force.

Oncotic pressure (colloid osmotic pressure) is the osmotic pressure resulting from the difference within the ECF between the protein contents of plasma and interstitial fluid. The major contribution to oncotic pressure under physiological conditions is plasma albumin concentration. The hydrostatic pressure of the plasma across the afferent capillary membrane creates a counter force to the oncotic pressure: the combination of the changing hydrostatic and oncotic pressure gradients across the capillary bed is known as the Starling forces.

Because the total solute content of cells under physiological conditions is essentially fixed and water is freely permeable across most cell membranes, the volume of the ICF is determined by the body water content. Intracellular fluid tonicity will in turn determine ECF tonicity, but ECF volume is essentially dependent on its sodium content.

Extracellular fluid and sodium

The sodium content of a normal adult is 55–65 mmol/kg body weight. The concentration of sodium in plasma is ~140 mmol/L (~152 mmol/kg). Under physiological conditions, the control of the ECF volume is through the control of functioning or *effective* plasma volume (that part of the plasma volume actively perfusing tissues). There are a variety of afferent mechanisms to monitor effective plasma volume (and thus ECF volume), which include intrathoracic volume receptors such as atrial stretch receptors, hepatic volume receptors, arterial baroreceptors, intrarenal baroreceptors and, possibly, tissue receptors monitoring tissue perfusion. Whatever the actual or relative function of all these sensory systems, their resultant influence is fine control over the renal conservation of sodium and the appetite for oral sodium intake.

Sodium intake varies considerably between different human cultural and ethnic groups. Variations in intake between 5 and 500 mmol/24h have been recorded and physiological mechanisms balance this with the renal excretion of sodium. Sodium delivery to the renal tubules is a function of plasma sodium concentration and the glomerular filtration rate (GFR). Every 24h, the kidneys of an average healthy adult male will filter in excess of 24 000 mmol of sodium, most of which is reabsorbed in the tubules so that, in health, sodium balance is achieved. In a healthy individual, renal sodium conservation can be extremely efficient, with urine sodium concentration falling to <1 mmol/L urine. Conversely, when sodium intake is excessive, the capacity to excrete sodium can result in urine sodium concentrations up to 300 mmol/L urine.

Renal control of sodium output

Intrinsic renal control of tubular reabsorption of sodium. Under normal physiological conditions, approximately 80% of the sodium in the glomerular filtrate is reabsorbed in the proximal tubules. The protein concentration of the blood within the post-glomerular peritubular capillary beds is believed to exert a strong oncotic pressure on fluid in the proximal tubules, and this in turn helps to regulate the volume of fluid reabsorbed. This process contributes to the autoregulation of filtration and reabsorption known as *glomerulo–tubular balance*. There has been considerable physiological interest in the control of proximal tubular sodium reabsorption and other intrinsic renal control mechanisms, such as redistribution of filtering activity from superficial nephrons (relatively salt-losing) to juxtamedullary nephrons (relatively salt-retaining). However, to date, the major humoral influences on sodium reabsorption appear to reside in the distal tubules and collecting ducts.

Renin–angiotensin–aldosterone axis. Aldosterone is a steroid hormone released from the zona glomerulosa of the adrenal cortex. The major control of aldosterone secretion is through angiotensin II, an octapeptide produced in the circulation as a final product of the action of renin. Renin is a proteolytic enzyme secreted by a group of

cells (the juxtaglomerular apparatus) situated between the afferent and efferent glomerular arterioles, and specialist chemoreceptor cells found within the distal convoluted tubular epithelium of the kidneys – the macula densa. Renin release is stimulated by decreased sodium delivery to the distal tubules (sodium depletion, ECF volume contraction), renal artery hypotension (due to systemic hypotension or renal artery stenosis) and by sympathetic nerve activity via β_1 -adrenergic receptors. The substrate for renin is angiotensinogen, an α_2 -globulin produced by the liver. Renin releases an amino-terminal decapeptide from angiotensinogen known as angiotensin I, which in turn is acted upon by angiotensin-converting enzyme (ACE), predominantly within the pulmonary capillaries. The action of ACE is to cleave the carboxy-terminal dipeptide of angiotensin I to produce angiotensin II. The renin–angiotensin–aldosterone axis is summarized in Figure 4.1. The major physiological influences on aldosterone are body sodium content and renal perfusion pressure, although hyperkalaemia can stimulate aldosterone release directly.

Aldosterone acts through the specific nuclear mineralocorticoid receptor, which is protected from cortisol (for which it has equal affinity) through the intracellular production of 11β -hydroxysteroid dehydrogenase (11β -HSD). This enzyme converts cortisol to cortisone, for which the receptor has weak affinity. The response within the *principal cells* lining the distal tubules and collecting ducts is the apical influx of sodium via epithelial Na^+ channel (ENaC) stimulation and its efflux via basolateral

Na^+, K^+ -ATPase. The net effect is active sodium reabsorption in exchange for potassium. In addition, angiotensin II has direct vasoconstrictive actions, thus having an immediate influence on *effective* plasma volume.

Natriuretic peptides. Glomerular filtration and the action of aldosterone do not constitute the complete control over renal sodium excretion within mammalian systems. The existence of a third factor (or factors) had been proposed for 20 years prior to the identification of a specific natriuretic factor in 1981. This factor was originally identified in rat cardiac atria and was termed atrial natriuretic factor (now known as atrial natriuretic peptide, ANP). Since then, a further two natriuretic peptides have been identified in humans. Circulating ANP is a 28 amino acid (AA) peptide with a 17 AA ring structure formed by a disulphide bridge between cysteines at positions 7 and 23: the gene for the ANP precursor molecule is located on the short arm of chromosome 1. Three exons code for a 151 AA peptide (preproANP) which, following removal of the signal peptide, results in a 126 AA peptide (proANP) – the main storage form. On secretion into the circulation, proANP is cleaved into the N-terminal 1–98 peptide (NT-proANP), and the biologically active 99–126 peptide (ANP). The major stimulus to the secretion of ANP is atrial stretch and the major sites of synthesis are the atria.

In 1988, a second natriuretic peptide was identified, in porcine brain, and termed brain natriuretic peptide (BNP). It was later shown to be produced predominantly in the ventricles of the heart. Circulating BNP is a 32

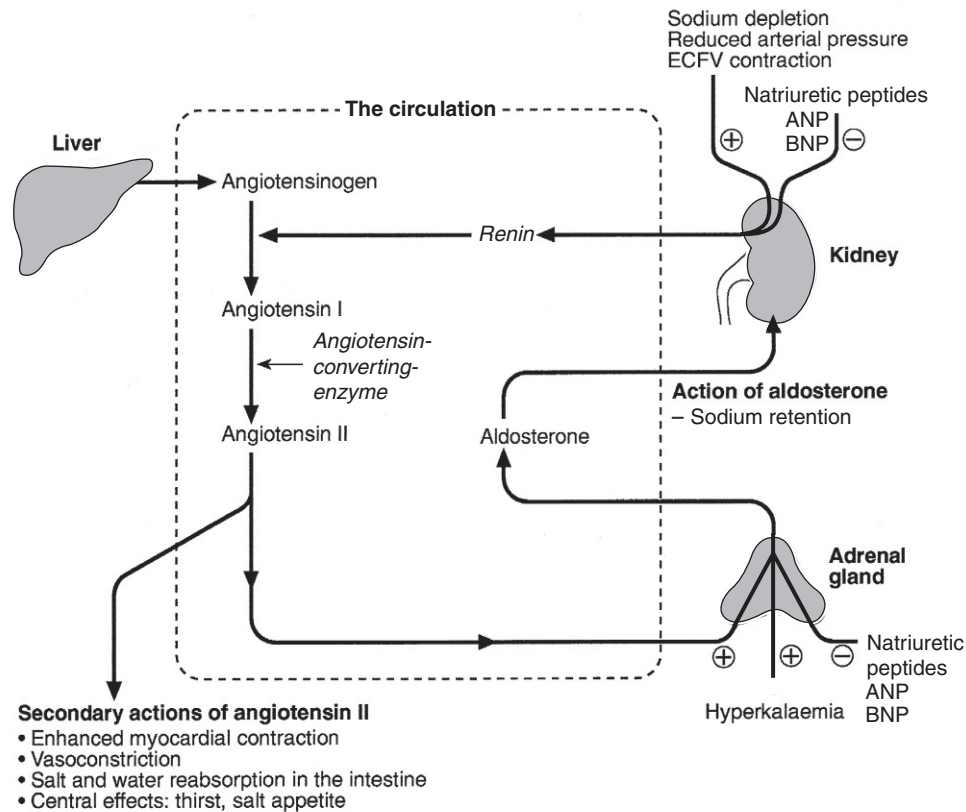


FIGURE 4.1 ■ The renin–angiotensin–aldosterone system. +, stimulatory signal; –, inhibitory signal; ANP, atrial natriuretic peptide; BNP, brain natriuretic peptide.

AA peptide, also with a 17 AA ring structure formed by a disulphide bridge between cysteines at positions 10 and 26. The ring structure has a high level of homology with the ring of ANP. Also like that of ANP, the gene for BNP is located on the short arm of chromosome 1. Three exons code for a 134 AA peptide (pre-proBNP) which, following removal of the signal peptide, results in a 108 AA peptide (proBNP). Further cleavage on release into the circulation results in the N-terminal 1–76 peptide (NT-proBNP) and the biologically active 77–108 peptide (BNP). Both ANP and BNP share a common receptor that mediates a natriuretic response in the kidneys by causing an increase in glomerular filtration rate and blocking sodium reabsorption in the inner medullary collecting ducts. They also influence sodium reabsorption through antagonism of the renin–angiotensin–aldosterone axis. Both ANP and BNP reduce sympathetic tone in the peripheral vasculature. The basic physiology of ANP and BNP is summarized in [Figure 4.2](#).

In 1990, a further ‘natriuretic peptide’ was described in porcine brain and termed C-type natriuretic peptide (CNP). The gene for CNP is coded on chromosome 4. Although CNP has been identified in human plasma, this peptide is mediated through a different receptor and does not have direct natriuretic function, but acts primarily as an antiproliferative regulator in the vascular cell system and as a neuropeptide.

There is no doubt that ANP, and to a lesser extent BNP, has important physiological functions in relation to sodium balance. Simple experiments can readily demonstrate changes in ANP and BNP concentrations in relation to dietary sodium intake, and the interaction of ANP and BNP with the renin–angiotensin–aldosterone axis results in a dual ‘fine-tuning’ system of sodium control, using afferent information obtained from the heart and kidneys simultaneously. However, unlike the situation with the renin–angiotensin–aldosterone axis, no primary disorders of natriuretic hormone excess or deficiency have yet been identified with certainty. Because the major stimulus to the release of ANP and BNP is cardiac wall

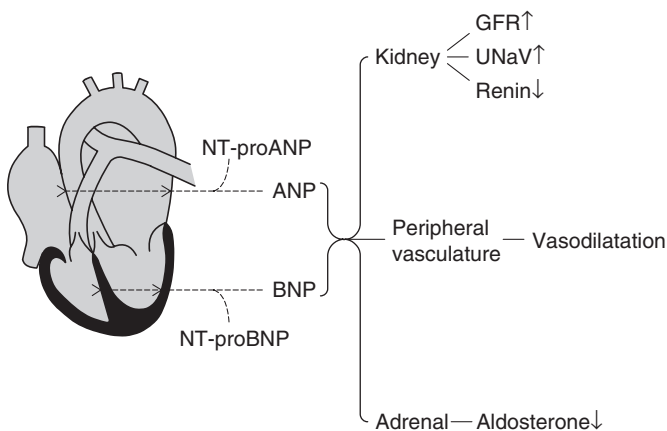


FIGURE 4.2 ■ The physiology of ANP and BNP. N-terminal propeptides are co-secreted with the active natriuretic peptides. Atrial natriuretic peptide (ANP) and brain natriuretic peptide (BNP) share a common receptor to increase glomerular filtration rate (GFR) and urine sodium excretion (UNaV), and to reduce renin and aldosterone secretion.

stretch, the major current clinical utility of measuring these peptides, particularly BNP and NT-proBNP, is in the diagnosis and monitoring of cardiac failure.

Sodium appetite

The renal conservation of sodium is remarkably efficient, but when sustained, non-renal losses occur under physiological conditions, such as through sweating due to prolonged vigorous physical activity or as a result of prolonged exposure to high ambient temperature, then a mechanism for increasing sodium intake comes into play – the salt appetite. That such a mechanism exists in humans can be observed in pathological states of impaired sodium conservation such as Addison disease. However, accurately defining the sodium appetite in humans under physiological conditions is a subjective and difficult task. Salt intakes are largely conditioned by traditions of food intake and cultural habits of seasoning food with salt. The impulse to add salt to prepared food appears to be automatic in many people, who do so even before tasting. Animal experiments have implicated an entirely separate, brain-based, renin–angiotensin–aldosterone system as a controlling influence on active salt-seeking behaviour.

Intracellular fluid and water

Water crosses cell membranes by simple diffusion through the lipid bilayer and through specific water channels known as aquaporins. The existence of specific water channels had been suspected for many years, not least to explain the dramatic changes that can occur in water permeability in the epithelial cells of the renal collecting ducts during hydration or dehydration. However, it was not until 1992 that the first aquaporin was characterized in human red cells – aquaporin 1 (AQP 1). To date, 13 different mammalian aquaporins have been identified (AQP 0–12). Specific human disease has been associated with mutations in the genes coding for AQP 0 (congenital cataracts) and AQP 2 (autosomal recessive nephrogenic diabetes insipidus – see [p. 41](#)). Although mutations in the gene for AQP 1 have been described, affected individuals remain asymptomatic. In addition, the condition neuromyelitis optica (NMO), a rare demyelinating and inflammatory disease of the central nervous system (CNS), is caused by autoantibodies directed against the main water channel of the CNS – AQP 4. The regulation of the gene expression of aquaporins may have important pathophysiological consequences in disease states associated with water retention or cerebral oedema.

Under physiological conditions, the solute content of cells is constant and, therefore, cell volume is dependent on solvent, not solute, content. The majority of cells behave as ‘effective’ osmotic meters – swelling when body water increases and contracting when body water decreases. Normally, the ECF and hence ICF osmolality are maintained at about 285 mmol/kg.

Control of renal water output

Osmoregulation. There is a minimum obligatory loss of water by the kidneys each day that is dependent upon

the maximum achievable urine concentration and the osmotic load for excretion. The maximum renal loss of water occurs when, for a given osmotic load, the minimum urine concentration is achieved.

The control of water output by the body is through secretion of antidiuretic hormone (arginine vasopressin, AVP) and its renal action. Arginine vasopressin is a nonapeptide synthesized in magnicellular neurons within two paired nuclei in the hypothalamus – the supraoptic and paraventricular nuclei. The gene for AVP in humans is located on the short arm of chromosome 20. Three exons code for a pre-pro-vasopressin that, following removal of the signal peptide, results in pro-vasopressin, which is subsequently packaged into neurosecretory granules. The granules are then transported by axonal flow to nerve terminals in the posterior pituitary. During transport, pro-vasopressin is further cleaved to AVP, neurophysin II and a glycoprotein. The neurophysin II forms tetramers, with one AVP molecule bound to each neurophysin moiety and the whole possessing a further AVP binding site. The stimulus to release AVP into the circulation results in the simultaneous release of AVP, neurophysin II and the glycoprotein. Closely associated cells within the hypothalamus (the osmoreceptor cells), by virtue of their swelling or shrinking in response to changes in ECF osmolality, control the release of AVP from the posterior pituitary. The effect of changing plasma sodium concentration (and hence osmolality) on plasma AVP concentration is shown in Figure 4.3. Other solutes confined to the ECF, for example exogenously administered mannitol, have a similar effect. By contrast, urea produces no significant stimulation of AVP because it freely permeates cell membranes. The osmoreceptor response is often characterized by its set point (variously defined by the plasma osmolality at which a measurable AVP response commences or by the basal state osmolality) and by its responsiveness (gain or sensitivity as judged by the slope of the response).

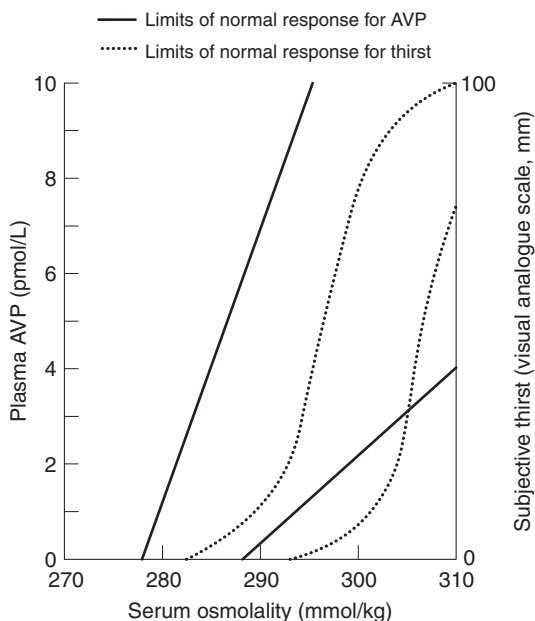


FIGURE 4.3 ■ Osmoregulation of arginine vasopressin (AVP) and thirst.

Thus, in situations of water depletion, ECF osmolality will increase, osmoreceptor cells will contract and plasma AVP secretion will increase. Hydration will reverse these events and suppress AVP. This system constitutes the osmoregulatory control of AVP release.

Non-osmotic control of arginine vasopressin. In addition to osmoregulation, certain non-osmotic controls over the secretion of AVP exist, including ECF hypovolaemia, hypotension, nausea and an oropharyngeal response. The AVP response to hypovolaemia and hypotension is relatively insensitive when changes are proportionally small (5–10% reductions), but increases exponentially as further reductions occur. Thus, a reduction in ECF volume or blood pressure of 20% or more will result in a plasma AVP concentration far in excess of that observed during normal osmoregulation. The influence of baroreceptor or volume receptor afferent input appears to modulate the osmotic response but does not abolish it: modulation occurs by decreasing the threshold for AVP release and increasing the gain of the systems.

Nausea is the single most powerful stimulus to AVP secretion. It overrides osmoregulatory control, and plasma AVP concentrations may increase 100-fold or more. A short-lived oropharyngeal suppression of AVP can occur following oral ingestion of fluid and prior to any reduction in serum osmolality, thus direct studies of osmoregulation of AVP need to avoid any such stimulus.

Renal responsiveness to arginine vasopressin. The apical (luminal) surface of the specialized cells lining the collecting ducts is essentially impermeable to water, except when AVP occupies its specific receptors on the basolateral (contraluminal) surface – the V_2 receptor (AVPR2). Stimulation of AVPR2 results in AQP 2 channels, located in vesicles beneath the apical membrane, fusing with that membrane and rapidly delivering water channels to the cell surface. A schematic representation of this model is shown in Figure 4.4. Under these conditions, water is absorbed into the collecting duct cells as a result of medullary hyperosmolality and is then absorbed into the bloodstream: concentrated urine is formed. During states of overhydration, when AVP secretion is suppressed, the luminal surfaces of the collecting duct cells remain impermeable to water. Luminal fluid rendered hypotonic in the diluting segments of the nephrons is not exposed to the medullary hypertonicity and, hence, no water is absorbed into the bloodstream; dilute urine is formed. The renal response to AVP is thus dependent on an intact receptor–effector mechanism within the collecting duct cells, resulting in an alteration in luminal membrane permeability, and upon the presence of a renal medullary osmotic gradient. Human urine osmolality varies between approximately 50 and 1400 mmol/kg.

Control of water intake

Osmoregulation. The physiological stimulus to water intake is thirst. However, the act of drinking in human societies in temperate regions is predominantly a social

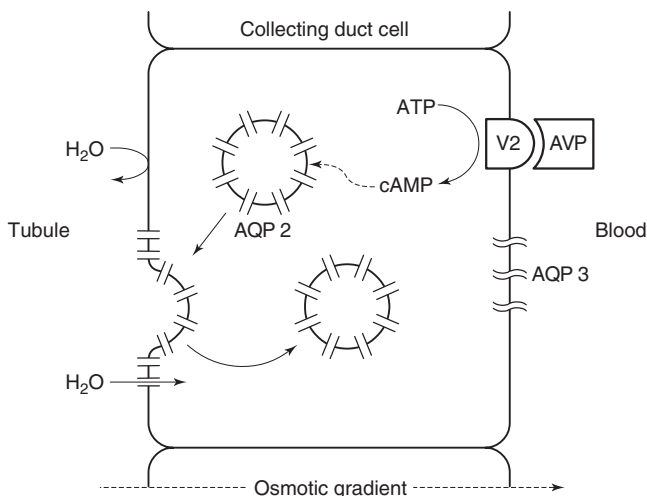


FIGURE 4.4 ■ Renal action of arginine vasopressin (AVP). Interaction of AVP with its V_2 receptor located on the basolateral membrane results in the generation of cAMP, which subsequently causes the fusion of AQP 2 laden vesicles with the otherwise impermeable apical membrane. Under the influence of the medullary osmotic gradient, water is drawn across the cell and is taken up within the blood via AVP-independent AQP 3 water channels.

or habitual act not dependent on thirst. The control of water balance under non-pathological conditions is thus, in many individuals and for much of the time, achieved by control of output.

The osmoregulation of water output by the kidneys can only modulate and restrict an established deficit, but to reverse such a deficit, a specific water input mechanism – thirst – is required. The osmoregulation of thirst is similar in principle to that of AVP and appears to be primarily governed by alteration in osmoreceptor cell size. The osmoreceptor cells controlling thirst are considered to be closely linked to, but distinct from, those controlling AVP secretion. A common method of study in human subjects is the measurement of subjective thirst during continuous osmotic stimulation with hypertonic saline infusion, and a typical range of normal responses is shown in Figure 4.3. Controversy exists as to the exact relationship between the onset of effective renal water conservation and the onset of effective thirst – that is, a sensation of thirst sufficient to cause active seeking and intake of water. However, physiologically, the onset of AVP secretion (and hence water conservation) differs from the onset of thirst and normally precedes it. The magnitude of this difference in water output and input effector controls will determine whether an individual controls physiological water balance primarily by thirst and water intake or by renal conservation of water.

Non-osmotic control of thirst. Non-osmotic thirst occurs when extracellular fluid is lost without corresponding cellular dehydration, the osmotic pressure of the extracellular fluid remaining unchanged. In this respect, the overall control of thirst parallels the control of AVP, with both osmotic and hypovolaemic stimuli. There is good evidence from animal experiments of both neural

and hormonal mediators controlling non-osmotic thirst. Angiotensin II is the most potent human thirst stimulant and may act directly upon the brain, but even when the effects of angiotensin II are blocked, significant hypovolaemia will still stimulate thirst.

Thirst following haemorrhage is a commonly reported clinical observation but, like the AVP responses to extracellular hypovolaemia, often a considerable degree of haemorrhage (15–20% of total blood volume) is necessary before the sensation becomes strong. Thus for day-to-day water balance, the primary physiological control of thirst is osmotic.

Extracellular fluid, intracellular fluid and potassium

In health, plasma potassium concentrations range between 3.1 and 4.6 mmol/L, with values in serum approximately 0.3–0.4 mmol/L greater, owing to release of potassium during clot formation (mean serum concentration 4.0 mmol/L). The intracellular concentration of potassium is ~160 mmol/kg, and 98% of the total body potassium is present in the intracellular fluid. There are two aspects to the physiological control of potassium, namely, the total body content and its distribution between intra- and extracellular spaces.

Extracellular and intracellular fluid distribution of potassium

The potassium content of cells is determined by the balance of activity between uptake of potassium due to membrane-bound Na^+, K^+ -ATPase and the passive loss or leakage of potassium out of the cell. Many factors can influence the distribution of potassium, for example acid–base status, hormones (insulin, catecholamines), osmolality and the cellular content of potassium. The influence of acid–base status is widely recognized as an important contributor to potassium distribution, with an association between hypokalaemia and alkalosis and between hyperkalaemia and acidosis, particularly when the acidosis is induced by mineral rather than organic acids.

Insulin promotes active uptake of potassium by cells, probably by direct stimulation of Na^+, K^+ -ATPase, and this activity appears to be independent of the effect of insulin on glucose uptake. The importance of the effects of insulin in controlling plasma potassium under physiological conditions is not understood, but its action has an important therapeutic role in the treatment of hyperkalaemia.

Catecholamines have an effect on potassium distribution, with β -adrenergic actions essentially promoting cellular uptake and α -adrenergic actions resulting in increases in plasma potassium concentration, but again the significance of these effects under physiological conditions is not understood. The net effect of catecholamines on cellular uptake of potassium probably explains the transient hypokalaemia frequently observed in acutely ill patients.

An acute increase in extracellular tonicity, such as occurs following hyperosmotic infusions of saline or mannitol, results in an increase in plasma potassium

concentration. This results from leakage of potassium from cells, a phenomenon that is not related to extracellular acidosis, but may be linked to cellular dehydration, altered cell membrane function or altered cell metabolism. An increase in extracellular tonicity is also observed in patients with hyperglycaemia in the absence of insulin and has important therapeutic relevance in the provision of potassium replacement during the treatment of hyperglycaemia. The effects of tonicity under physiological conditions are probably of no significance.

Potassium depletion results in a greater loss of potassium from the ECF than the ICF, and potassium excess results in a greater proportional rise of ECF than of ICF potassium concentration. The controlling influences over these changes are not defined, but the result is a significant alteration in membrane potential: this is increased with potassium depletion and decreased with excess. The effects on neuromuscular function of either condition constitute the most important clinical complications of disorders of potassium metabolism.

Renal control of potassium output

Intrinsic tubular control. The traditional understanding of potassium handling by the kidneys is that potassium is freely filtered by the glomeruli, but that up to 95% has been reabsorbed before the tubular fluid reaches the distal convoluted tubules. The predominant control of potassium excretion appears to reside in the control of distal tubular reabsorption and secretion.

Plasma potassium itself has a major effect on potassium secretion in the distal tubules, tending to correct any imbalance. Acute changes in sodium delivery to the distal tubules may also influence potassium excretion – restricted sodium delivery impairs potassium excretion, but a tendency to natriuresis is accompanied by a kaliuresis. However, chronic effects on potassium excretion as a result of changes in sodium intake are not seen because of the influence of the renin–angiotensin–aldosterone axis.

Aldosterone. Potassium directly influences aldosterone secretion from the adrenal cortex. A high plasma potassium concentration stimulates aldosterone secretion and a low concentration suppresses secretion. In addition to its effects on sodium reabsorption by the principal cells, aldosterone stimulates hydrogen ion secretion by the α -intercalated cells of the distal tubules and collecting ducts. Acidosis is associated with reduced potassium secretion and alkalosis with enhanced secretion. The net effect of aldosterone is to stimulate exchange of potassium and hydrogen ions for sodium ions. Therefore, the relative proportions of potassium and hydrogen ions within the cells of the distal tubules, together with the ability to secrete hydrogen ions, will determine the effect of systemic acidosis or alkalosis on potassium excretion. Acting alone, an acidosis will promote potassium retention and an alkalosis will promote a kaliuresis.

Urine potassium concentration can vary between about 5 mmol/L and 150 mmol/L. Adaptation of urinary excretion to a variation in input tends to be slow, taking a

few days to achieve a new balance. In this respect, urinary control of potassium is less sensitive than the control of sodium.

DISORDERS OF SODIUM METABOLISM

As sodium is predominantly an extracellular cation, the control of sodium balance will control the volume of the ECF. The tonicity of body fluids is under osmoregulatory control, therefore sodium deficit or sodium excess presents clinically with primary changes in ECF volume rather than changes in sodium concentration within the ECF. Hyponatraemic and hypernatraemic states are discussed in the section on water metabolism.

Sodium deficiency

Clinical presentation

Sodium is always lost from the body in association with water. As the sodium concentration of all body fluids is equal to or less than plasma (except on occasions of high sodium intake, when urine sodium concentration may exceed plasma concentration), loss of any body fluid except plasma will generally result in an excess loss of water over sodium. Any loss of sodium, however, will result in a reduction of ECF volume, including a reduction in circulating plasma volume. Clinical presentation will depend on the severity of the decrease. When the changes are mild, patients are often described clinically as being dehydrated, a description that should be confined to pure water deficiency but, unfortunately, in general usage is not. Except in rare instances, truly dehydrated (water depleted) patients are extremely thirsty; patients with all but the most severe salt and water deficiency are not.

Reduced intravascular volume, when mild, will result in postural hypotension and a compensatory increase in pulse rate; central venous pressure is reduced and this can be assessed clinically by observation of neck vein filling or directly measured following insertion of a cannula into a central vein. When the volume reduction is more severe, hypotension and eventually shock will result in oliguria; the central venous pressure is further reduced. Reduction in interstitial fluid volume results in reduced skin turgor and transcellular fluid reductions result in dry mouth and reduced intraocular pressure.

Causes of sodium deficiency

The causes of sodium deficiency can be classified broadly into extrarenal, primary renal (resulting from renal disease) and secondary renal (resulting from disturbed hormonal control of renal sodium retention or from inappropriate use or abuse of diuretics). In addition, and somewhat difficult to classify, is the sodium deficiency that can occur when isolated jejunal segments are incorporated into urinary diversion operations (jejunal conduit, jejunal continent diversion, see p. 55). These procedures are now rarely performed, as alternative sources of donor intestine are preferred. When jejunum is used in such procedures, there is a risk of postrenal loss of sodium.

Extrarenal sodium loss. Extrarenal fluids have sodium concentrations that may approach the concentration in plasma (Table 4.3). The major causes of extrarenal sodium deficiency are summarized in Box 4.1. The commonest clinical presentations result from gastrointestinal disease. The clinical history may considerably underplay the degree of deficit, especially in chronic conditions or conditions resulting in sequestration of fluid.

Primary renal sodium loss. The major causes of primary renal sodium loss are summarized in Box 4.2. The recovery phase of acute kidney injury is often associated with a polyuria, kaliuresis and natriuresis. Usually this stage is short lived, lasting only a few days, but it may occasionally be prolonged. A natriuresis may occur following successful renal transplantation and this may be in part due to transient tubular dysfunction; the recovery phase usually lasts only a few days, but occasional patients show prolonged natriuresis.

Relief of urinary tract obstruction, most commonly seen in patients with prostatic enlargement, is often followed by a short period of diuresis and natriuresis. The exact mechanism of this is not fully understood, but is probably related to a urea-induced osmotic diuresis, together with elimination of excess sodium retained during the obstruction phase. The natriuresis in these cases is corrective and is not strictly a primary renal condition: it is unlikely to lead to sodium depletion. Normally, this phase of post-obstructive natriuresis and diuresis lasts between one and seven days, but occasionally a more prolonged natriuresis occurs, leading to sodium deficiency. This rare event is secondary to tubular damage occurring during obstruction. Salt wasting has been described in association with meticcillin-induced acute interstitial nephritis. Full resolution of the condition may be delayed for several months, during which time minimum obligatory losses of sodium may exceed 100 mmol/24 h.

Chronic kidney disease is generally associated with a reduced capacity to excrete sodium. Considerable adaptation occurs to increase the natriuretic capacity of the remaining functioning nephrons, which paradoxically impairs the overall sodium conserving function. On a Western type diet, sodium intake normally exceeds minimum obligatory urine sodium loss, but if dietary salt is restricted or non-renal losses of sodium are increased, the obligatory loss may induce sodium deficiency. Minimum obligatory losses may be as little as 40 mmol/24 h and even lower if enough time is given for adaptation, but occasionally patients have minimum

obligatory losses exceeding 150 mmol/24 h (3 mmol/kg body weight in children). The term *salt-losing nephropathy* is appropriate for this group of patients.

Salt-losing nephropathy is a clinical state, rather than a specific disease. It is normally associated with chronic kidney disease due to tubulointerstitial disease or to glomerulonephritis with significant interstitial abnormalities. A common example of a toxic nephropathy inducing a salt-losing state is analgesic abuse, in which there is typically also reduced concentrating ability and renal tubular acidosis.

BOX 4.1 Causes of extrarenal sodium loss

Gastrointestinal

- Gastric, e.g. vomiting, aspiration, fistula, blood loss
- Midgut, e.g. ileostomy, pancreatic and biliary fistulae
- Hindgut, e.g. diarrhoea, bleeding, excess mucus production

Skin

- Sweating, e.g. thermal or increased sodium content (cystic fibrosis)
- Burns
- Exudative skin disease

Sequestration

- Gastrointestinal, e.g. ileus, small bowel obstruction
- Other transcellular spaces, e.g. peritonitis, pleural effusion

BOX 4.2 Causes of primary renal sodium loss

Acute

- Diuretic phase of acute kidney disease
- Postrenal transplantation
- Following relief of urinary tract obstruction
- Acute interstitial nephritis

Chronic

- Chronic kidney disease with salt restriction
- Salt-losing nephropathy
 - Chronic pyelonephritis
 - Medullary cystic disease
 - Toxic nephropathies, e.g. chronic analgesic abuse, cisplatin

TABLE 4.3 Examples of fluid composition of extrarenal fluids

	Sodium (mmol/L)	Potassium (mmol/L)	Chloride (mmol/L)	Bicarbonate (mmol/L)
Gastric juice (with parietal cell secretion)	20–70	5–15	80–160	0
Pancreatic juice	140	6–9	110–130	25–45
Bile	130–165	3–12	90–120	30
Ileal fluid	105–144	6–29	42–60	50–70
Stool water	32–40	75–90	12–18	30–40
Sweat	5–80	5–15	5–70	–

Secondary renal sodium loss. The disruption of extrinsic controls over renal sodium handling can give rise to a secondary renal sodium loss. Patients present with a variety of symptoms, some specifically related to a contracted ECF volume, such as postural hypotension and an increased sodium appetite. Causes of secondary renal sodium loss, presented in [Box 4.3](#), are essentially hormone or diuretic induced. In Addison disease, the synthesis and secretion of aldosterone are reduced because of adrenal gland destruction, leading to distal renal tubular sodium wasting. Congenital adrenal hyperplasia (CAH) with associated sodium wasting is a result of impaired mineralocorticoid synthesis, most commonly as a consequence of 21-hydroxylase deficiency, but also, less commonly, 3 β -hydroxysteroid dehydrogenase deficiency. In either condition, the degree of sodium wasting is variable: it is present in about two-thirds of patients with 21-hydroxylase deficiency (see Chapter 18). Very rare forms also exist, including cholesterol desmolase deficiency (lipoid adrenal hyperplasia) and corticosterone methyl oxidase deficiency – a deficiency in the mixed-function oxidase catalysing the final steps of aldosterone synthesis. Corticosterone methyl oxidase deficiency is not strictly a type of CAH, as the synthesis of cortisol is unaffected and, consequently, the adrenals are not hyperplastic. In all these conditions, the associated hypovolaemia stimulates renin production, and all may be associated with hyperkalaemia. Treatment is primarily by glucocorticoid and mineralocorticoid replacement.

It might be predicted that deficient production of renin would also lead to renal sodium loss because of a secondary deficiency of aldosterone. The condition hyporeninaemic hypoaldosteronism is invariably associated with renal insufficiency, but the characteristic feature is hyperkalaemia rather than hypovolaemia. As in the majority of patients with renal insufficiency, inappropriate renal sodium loss will induce hypovolaemia if the patient is placed on a low sodium intake, but it is unclear if this is more severe in hyporeninaemic hypoaldosteronism. Although this condition is a cause of secondary renal sodium loss, its main importance is in the differential diagnosis of hyperkalaemia and it is discussed in more detail later in this chapter (see [Syndromes of hypoaldosteronism, p. 60](#)).

Pseudohypoaldosteronism (type 1) is a rare congenital disease existing in two forms, caused by an end organ failure of response to aldosterone within the principal cells

BOX 4.3 Causes of secondary renal salt loss

- Hyper-reninaemic hypoaldosteronism
 - Addison disease
 - Congenital adrenal hyperplasia
 - Corticosterone methyl oxidase deficiency types I and II
- Hyporeninaemic hypoaldosteronism
- Pseudohypoaldosteronism (type 1)
- Diuretics
- Bartter syndrome
- Gitelman syndrome

lining the renal collecting ducts. It is caused by either a loss of function mutation in the mineralocorticoid receptor (MR) gene (autosomal dominant form) or a mutation resulting in loss of function in the epithelial Na⁺ channel (ENaC – autosomal recessive form). Infants present with dehydration, hyponatraemia, hyperkalaemia, metabolic acidosis, failure to thrive and weight loss. Renal sodium wasting is unresponsive to exogenous mineralocorticoids. Pseudohypoaldosteronism is also discussed in more detail later in this chapter (see [Syndromes of hypoaldosteronism, p. 60](#)).

Diuretic abuse, especially when surreptitious, may present a difficult diagnostic problem. The major potent diuretics, such as furosemide or bumetanide, may induce significant renal sodium loss and the differential diagnosis includes Bartter syndrome. Chronic abuse with thiazide diuretics may produce a clinical picture similar to Gitelman syndrome. In all cases, however, the major and persistent finding is significant hypokalaemia (see [p. 55](#)).

Laboratory investigation of sodium deficiency

No single laboratory finding is diagnostic of sodium deficiency, which, therefore, remains primarily a clinical diagnosis. There may be uraemia, particularly of a prerenal pattern (with urea concentration disproportionately elevated in relation to that of creatinine). Serum sodium concentration may be normal, decreased (usually in association with uraemia) or even increased when the sodium deficit is proportionately less than the water deficit. Serum sodium and urea concentrations cannot indicate accurately the degree of sodium depletion or reduction in ECF volume.

Measurement of the haematocrit may help to establish the degree of reduction in ECF volume, especially if the individual's normal haematocrit is known (see [Appendix 4.1a](#)). Similarly, serum protein measurements may suggest ECF volume contraction. In a steady state, hyponatraemia in association with clinical evidence of sodium deficit can also be used to estimate the extent of sodium deficit (see [Appendix 4.1b](#)). Either method provides only a rough guide to subsequent replacement.

Urine electrolytes will help to establish whether the loss has been renal or extrarenal. If there has been gastrointestinal loss of sodium or another extrarenal loss, renal sodium retention should be maximized. Random urine sodium concentrations will show conservation of sodium with a concentration <10 mmol/L. The mature, fully functioning renal system can reduce urine sodium concentration to <1 mmol/L, but this capacity is less in the very young and old. Twenty-four hour urine sodium output will show significant sodium retention, with 24h urine sodium excretion as low as 1 mmol and invariably <15 mmol, even in the elderly. Serum osmolality measurements are of no value in the differential diagnosis of sodium depletion, adding nothing to the information obtained from serum electrolytes, urea and creatinine measurements. Urine osmolality will generally show an overall significant increase towards maximum, reflecting the oliguric state, especially when sodium depletion is sufficient to cause a non-osmotic increase in arginine vasopressin (AVP).

Only when a renal loss of sodium is present without evidence of renal impairment should secondary hormonal or diuretic-related sodium depletion be considered. For the majority of hypoaldosterone-related disorders, there will be associated hyperkalaemia (for specific investigation of hypoaldosteronism, see Chapter 18). Diuretic abuse can provide an intriguing puzzle as, along with Bartter syndrome and Gitelman syndrome, it provides the major differential diagnosis of renal sodium loss in association with hypokalaemia.

Management of sodium deficiency

The essential steps in treating sodium deficiency and ECF volume depletion are to attempt to treat the causes of sodium and water loss and to adequately replace the fluid already lost. The amount of fluid replacement must be balanced against measured or estimated losses, both intra- and extracorporeal. The measurement of body weight is a useful adjunct to monitoring; an increase in weight may indicate accumulation of interstitial or transcellular fluid. Accurate fluid balance charts (see Table 4.4) should indicate a positive or negative balance, but even the most assiduously prepared charts (many are not) provide only an approximate estimate of balance, especially when confounding problems such as pyrexia (increasing insensible losses) or hypermetabolic states (increasing metabolic water) intervene. Vital signs: pulse, blood pressure and central venous pressure (if measured), all provide additional useful information to improve the practice of the art.

The type of replacement will be dependent on cause and severity. Mild forms of sodium deficit, such

as caused by overtreatment with diuretics or chronic salt-losing nephropathy, may be adequately treated with oral sodium and water supplementation. For more severe forms of sodium deficit, intravenous infusion is required; the types of fluids available are shown in Table 4.5. All of them are iso- or hypoosmolal to normal plasma. Hyperosmolal fluids, such as 1.8% or 5% sodium chloride or 2.74% or 8.4% sodium bicarbonate, are not appropriate treatments for sodium depletion with clinical evidence of reduced ECF volume, even in the presence of hyponatraemia, as infusions of such fluids will increase ECF volume in part by shifting fluid from the ICF volume.

Hypoosmolal fluids may be used when sodium deficit is associated with a greater degree of water deficit, that is, sodium deficit in association with hypernatraemia.

Sodium excess

Clinical presentation

Sodium excess is almost invariably associated with water excess (and is usually an iso-osmolal sodium excess). The clinical presentation will depend on the severity of the expansion of ECF volume and its relative distribution within the ECF volume compartments. In practice, the clinical presentation is with oedema (peripheral or pulmonary) or effusions (pleural, ascites) or with hypertension.

Peripheral oedema, due to sodium retention, is characteristically 'pitting' when digital pressure is applied to the affected part (lymphoedema, due to lymphatic obstruction, is characteristically non-pitting, while in myxoedema, the swelling is brawny in nature). Measurement or clinical observation of the central venous pressure may suggest an associated increase in intravascular volume, as may the demonstration of hepatic congestion or the presence of pulmonary crepitations.

Hypertension is a common clinical finding, but hypertension specifically due to chronic iso-osmolal sodium excess is rare. Causes of this form of hypertension are discussed later, but all are associated with an increased mineralocorticoid effect in the initial stages, promoting renal sodium conservation and renal loss of potassium. The association of muscle weakness (caused by hypokalaemia) with hypertension is an important clinical pointer to such a cause of sodium retention.

TABLE 4.4 Ideal components of a fluid balance chart

Input	Output	Extravasation
Oral fluid	Urine output	Body weight
Intravenous fluid	Gastrointestinal losses	
Water generated in metabolism (~500 mL/24 h)	Surgical drains etc.	
	Insensible losses (~900 mL/24 h)	

TABLE 4.5 Iso- and hypoosmolal fluids for treatment of sodium and water deficiency

	Na ⁺ (mmol/L)	K ⁺ (mmol/L)	Cl ⁻ (mmol/L)	HCO ₃ ⁻ (mmol/L)	Glucose (mmol/L)	Ca ²⁺ (mmol/L)	Osmolality (mmol/kg)
0.9% saline	154	–	154	–	–	–	290
0.45% saline	77	–	77	–	–	–	145
0.18% saline	31	–	31	–	–	–	58
Dextrose saline (2.5/0.45%)	77	–	77	–	140	–	290
(4.0/0.18%)	31	–	31	–	222	–	290
Ringer's solution	147	4.2	156	–	–	2.2	290
Hartmann's solution	131	5.4	112	29	–	1.8	290
Sodium lactate	167	–	–	(≅167)	–	–	290

Causes of sodium excess

Sodium excess with oedema. The causes of sodium excess in association with oedema formation are shown in Box 4.4. Congestive cardiac failure (CCF) is a common clinical example of sodium excess, presenting with dependent oedema, congested liver, increased jugular venous pressure and pulmonary crepitations due to oedema. The 'backward failure' hypothesis suggests that increased transcapillary pressure resulting from increased central venous pressure causes oedema and hence reduces intravascular volume. Reduced intravascular volume will in turn stimulate sodium retention. Although readily understandable, this theory is inadequate because patients generally have increased intravascular volumes. The 'forward failure' hypothesis suggests inadequate perfusion of the kidneys because of reduced cardiac output, which in turn promotes sodium retention. However, high cardiac output failure demonstrates that sodium retention can be independent of cardiac output. The reduced 'effective circulating volume hypothesis' lacks precise definition, but suggests that only a proportion of the volume is 'effective', that is, circulating volume appears to be reduced in that the renin-angiotensin-aldosterone mechanism is activated, as is the release of AVP. The onset of CCF causes ANP and, particularly, BNP, release, leading to markedly elevated plasma concentrations. However, the natriuretic effect of the hormones is severely blunted and corrective natriuresis does not occur without therapeutic intervention, with diuretics or with exogenous natriuretic peptide administration in pharmacological doses.

The nephrotic syndrome is characterized by severe glomerular proteinuria, hypoalbuminaemia and oedema.

BOX 4.4 Causes of sodium excess

Sodium excess with oedema

- Congestive cardiac failure
- Nephrotic syndrome
- Liver disease
- Pregnancy
- Menstrual cycle
- Idiopathic oedema

Sodium excess without oedema

- Acute sodium loading (excess i.v. or oral saline)
- Renal sodium retention
 - Primary hyperaldosteronism: adenoma, idiopathic hyperplasia, carcinoma, glucocorticoid suppressible
 - Secondary hyperaldosteronism with hypertension: renin secreting tumour, renovascular hypertension, accelerated hypertension
 - Glucocorticoid excess (mineralocorticoid function): Cushing disease, ectopic adrenocorticotrophic hormone (ACTH) secretion, exogenous glucocorticoids
 - Exogenous mineralocorticoids: prescribed mineralocorticoids, glycyrrhizic acid (liquorice), glycyrrhetic acid (carbenoxolone)
 - Congenital adrenal hyperplasia: 11 β -hydroxylase deficiency, 17 α -hydroxylase deficiency
 - Pseudohyperaldosteronism: Liddle syndrome, apparent mineralocorticoid excess (cortisol 11 β -hydroxysteroid dehydrogenase deficiency)

The classic hypothesis explained the oedema and secondary sodium retention, on the basis of reduced oncotic pressure owing to hypoalbuminaemia. However, the rare condition congenital analbuminaemia is not associated with significant oedema, and nephrotic patients infused with albumin do not consistently respond with a natriuresis and reduction of oedema. Sudden remission of minimal-change glomerulonephritis may result in clinical resolution prior to an increase in plasma albumin, and the presumed reduced circulating volume in nephrotic syndrome is less commonly found when sought than normal or increased circulating volume. These observations imply an intrarenal mechanism of sodium retention in the nephrotic syndrome. Sodium retention does not appear to be primarily driven by the renin-angiotensin-aldosterone system; moreover, ANP concentrations are moderately increased, not decreased, and the natriuretic effect of the hormone is maintained.

Chronic liver disease frequently results in a state of sodium retention leading to oedema and ascites. Again, the physiology of sodium retention is multifactorial, resulting from decreases in glomerular filtration, enhanced proximal tubular reabsorption and increased renin-angiotensin-aldosterone production, presumably from decreased 'effective' intravascular volume. As in CCF, the plasma ANP concentration is significantly increased, but again the natriuretic response is blunted. The concentration of AVP may also be significantly increased.

Pregnancy. Pregnancy demands a retention of between 700 and 1000 mmol of sodium in order to expand the maternal plasma volume and interstitial space and to provide the fetus with sodium. Sodium retention occurs progressively during pregnancy, even though the glomerular filtration rate is greatly increased, so that the reabsorption of filtered sodium must be substantially increased above the non-pregnant level. The renin-angiotensin-aldosterone system is stimulated during pregnancy, a finding that supports the theory of renal sodium retention as a response to reduced 'effective' vascular volume (due to redistribution of ECF volume to the growing fetus and its maternal circulatory support). However, many studies have shown increased plasma ANP concentrations during pregnancy, a finding somewhat at odds with the presumed reduced 'effective' vascular volume. Thus, the kidneys may retain sodium in response to some other stimulus (such as the effect of oestrogens), and the increase in ANP may represent the response to the resultant increased atrial pressure from expanded vascular volume. However, these two apparently conflicting theories are not mutually exclusive and it seems likely that the resultant increase both in the activity of the renin-angiotensin-aldosterone system and the secretion of ANP forms the basis for an enhanced control over sodium balance. Enormous differences in daily sodium intake (from <10 mmol to >300 mmol) can be adequately accommodated during pregnancy.

Dependent leg oedema is extremely common during pregnancy, especially during the third trimester, and much of this is attributable to the mechanical pressure of the uterus upon the venous return from the lower limbs. However, there is a general tendency to an increased

interstitial fluid content because of reduced oncotic pressure (as a result of reduced plasma protein concentration) and because of the effects of oestrogens, which enhance the hydration of the mucopolysaccharide ground substance of connective tissue. This general tendency to increased interstitial fluid content may also produce a more generalized oedema.

Excessive sodium retention occurs in pre-eclampsia. Total body sodium is increased, but plasma volume is usually decreased, with shifts of fluid to the interstitium. The reduced plasma volume is the result of intense vasoconstriction, which also results in hypertension. In pre-eclampsia, renin activity and aldosterone concentrations in plasma in the third trimester are less than those observed during normal pregnancy. Furthermore, ANP concentrations are significantly increased above those found in normal pregnancy, or indeed those found in pregnancy associated with essential hypertension. Atrial natriuretic peptide appears to be released into the maternal circulation in response to vasoconstriction and an increased volume load to the cardiac atria. Thus in pre-eclampsia, as compared with normal pregnancy, renin, aldosterone and ANP all act as if to promote natriuresis, but, for reasons which are not understood, natriuresis is impaired.

Menstrual cycle. Despite intensive study over 50 years or so, the question of whether the normal menstrual cycle is associated with sodium retention remains unresolved. There is widespread belief that the majority of women retain sodium premenstrually, giving rise to premenstrual oedema in some. Plasma aldosterone and renin concentrations are elevated during the luteal phase of the ovulatory cycle, but this elevation is coupled with an increase in progesterone, which is thought to antagonize the renal actions of aldosterone. Studies have shown that plasma ANP concentrations remain the same during the follicular and luteal phases. It has also been demonstrated that, in healthy females, the ANP response to volume expansion is no different in the follicular and luteal phases of the menstrual cycle, and the proportional suppression of renin and aldosterone is also identical between the two phases of the cycle. When these findings are coupled with the observations that body weight, creatinine clearance and basal sodium excretion do not alter during the phases of the cycle, it is apparent that there is little evidence of renal sodium retention. This remains an area for further research, particularly as inappropriate treatment of perceived symptoms may lead to further complications of sodium balance (see below).

Idiopathic oedema. Idiopathic oedema, sometimes known as cyclical oedema, is a condition that occurs in females postpuberty. Oedema of the face, hands and legs can develop rapidly, and weight gains up to 4 kg in 24 h have been described. The aetiology of this condition is unknown, but one suggestion is that diuretics may be used to produce rapid weight loss and that cessation of this self-medication leads to rebound sodium retention. The diuretic is subsequently recommenced for cosmetic reasons to reduce oedema formation and weight gain. Thus, a vicious cycle of sodium depletion

followed by sodium retention and oedema is set in place. However, in some patients prior diuretic use cannot be demonstrated.

Many patients with idiopathic oedema show exaggerated sodium retention on assuming an upright posture with a marked reduction in renal blood flow and GFR. Oestrogens do not appear to play an important role, as the condition has been demonstrated in a patient who had previously undergone bilateral oophorectomy. The role of renin–angiotensin–aldosterone is not fully understood, with a proportion of patients having increased plasma renin and aldosterone concentrations, particularly when in an upright posture, but with suppression of concentrations by a high salt diet. Limited studies of ANP have been performed, but no abnormalities have been demonstrated: normal basal values are found and are stimulated by volume expansion.

Sodium excess without oedema. The causes of sodium excess without oedema are shown in [Box 4.4](#). Acute sodium loading is a rare event, invariably caused by inappropriate sodium administration to highly dependent individuals, for example those in intensive therapy units, especially the very young, the old and the highly incapacitated. Acute hypernatraemia is a powerful stimulus for thirst and thus, to maintain the condition, the intake of fluid must be prevented by an inability, for whatever reason, to express or act upon the desire to drink. Examples of acute sodium loading include administration of high sodium concentration oral feeds to infants (either accidentally or as a form of abuse), the use of oral hypertonic sodium chloride solutions as emetics, excessive administration of intravenous hypertonic sodium bicarbonate and the voluntary consumption of excessive quantities of table salt. In the past, the accidental introduction of hypertonic sodium into the circulation during therapeutic abortion has been a cause of acute sodium excess. Acute sodium loading associated with hypernatraemia is a life-threatening condition and should be treated promptly with free water administration.

Mineralocorticoid excess has a variety of causes (see [Box 4.4](#)). Although this is a condition of excess sodium retention, the usual clinical presentation is hypertension with hypokalaemia. The severity of hypokalaemia is in turn dependent on sodium intake, being reduced in severity if sodium intake is reduced. All of these conditions are discussed elsewhere in this book, either as part of the pathology of the adrenal glands (see Chapter 18) or in relation to hypokalaemia (see [p. 55](#)). The fascination of these conditions with respect to sodium is why sodium retention does not usually progress to increase the volume of all the ECF spaces sufficiently to cause oedema.

As the ECF volume expands, the filtered load of sodium increases and the fraction reabsorbed by the renal tubules decreases. This mechanism constitutes the ‘escape’ for sodium (but not potassium) from mineralocorticoid effects. The underlying mechanism of escape has been the subject of investigation since it was first described over 50 years ago. The mechanism appears multifactorial with a haemodynamic response to increased intravascular volume resulting in reduced proximal reabsorption of sodium. There is good evidence

that natriuretic peptides may account for a part of the escape mechanism – ANP concentrations are elevated in primary hyperaldosteronism and return to normal when the effects of hyperaldosteronism are reversed or antagonized. Evidence also exists of a decrease in the thiazide-sensitive $\text{Na}^+\text{-Cl}^-$ co-transporter (NCCT) within the distal convoluted tubules. In addition intrarenal substances such as prostaglandins, kinins and nitric oxide may also promote natriuresis.

Laboratory investigation of sodium excess

The investigation of sodium excess is confined almost exclusively to the investigation of oedema and of hypertension; hypernatraemia is not a feature of sodium excess except in rare instances of acute sodium loading, when the cause is usually clear from the clinical history.

In practice, the purpose of the investigation of oedema is primarily to differentiate reduced oncotic pressure from other causes, usually by the measurement of serum protein concentration and the confirmation of any route of protein loss, for example the measurement of urine protein excretion. The diagnosis of oedema secondary to CCF and liver disease is primarily clinical; detailed sodium balance studies are not normally justified or, indeed, practicable. Idiopathic oedema is also a condition that is initially diagnosed from the clinical history, although periods of continuous sodium loading with simultaneous sodium balance studies may reveal a propensity for sodium retention.

The laboratory investigation of sodium excess as a contributor to hypertension involves the demonstration of increased mineralocorticoid action, the simplest pointer being the association of hypokalaemia and hypertension. Investigation of primary hyperaldosteronism, conditions of increased glucocorticoid secretion and conditions of sodium retention due to congenital adrenal hyperplasia are covered in Chapter 18. Other causes of hypokalaemia and hypertension are discussed in the section on potassium metabolism (see p. 55).

Management of sodium excess

For those conditions associated with oedema, the management will be aimed primarily at ameliorating the primary cause. This primary treatment, when applicable, needs to be coupled with attempts to control sodium and water retention by the restriction of dietary sodium and, when appropriate, the use of diuretics. The laboratory's role in management is to monitor the potential complications of diuretic treatment, notably hypokalaemia (especially with the use of loop diuretics), hyperkalaemia (with the use of potassium-sparing diuretics such as spironolactone) and hyponatraemia (when over-diuresis results in non-osmotic stimulation of AVP with secondary water retention). In addition, the laboratory's function is to monitor the effects of natriuresis on renal function by monitoring serum urea or creatinine concentrations.

For conditions of sodium excess without oedema, the initial management is to increase renal sodium excretion,

either by attention to iatrogenic causes (such as exogenous mineralocorticoid treatment) or by treatment with diuretics such as spironolactone. This treatment is combined, when appropriate, with antihypertensive therapy. The laboratory's role in management is to monitor for possible complications, particularly with respect to potassium homeostasis. For those patients with surgically correctable conditions, such as primary hyperaldosteronism due to an adrenal adenoma, the definitive treatment is surgery.

DISORDERS OF WATER METABOLISM

As water is distributed throughout all body spaces, the effects of pathological excess or deficiency will be reflected in the relative functional sensitivities of each space. For example, a 10% excess or reduction in total body water (TBW) is unlikely to result in clinical features of alterations in ECF volume. However, ICF volume changes may considerably impair cellular function; in particular, rapid changes may significantly impair brain cell function.

Disorders of water metabolism include three categories of conditions. First are conditions in which polyuria is the major feature. Polyuria is defined as a urine output in adults in excess of 50 mL/kg body weight/24h. In polyuria, one aspect of water homeostasis is defective but may be compensated by another: either a reduced ability to concentrate the urine is compensated by a secondary increase in thirst, or an excessive water intake is compensated with a secondary increase in urine output. In either case, patients are usually normonatremic unless the compensatory mechanism is compromised.

Second are those of water deficiency in which homeostasis is defective and normal body water content cannot be maintained: patients present with hypernatraemia (serum sodium above 145 mmol/L).

Third are conditions of primary water excess in which homeostasis is defective: patients present with hyponatraemia (usually defined as a serum sodium <130 mmol/L) or clinical features of water intoxication.

Polyuria

Primary polyuria with secondary polydipsia

Polyuria is an early feature of diabetes mellitus as a result of the osmotic effects of an increased filtered load of glucose. Polyuria can also be a feature of renal failure due to the loss of medullary hypertonicity and the reduction in the production of AQP 2. Thus polyuria due to diabetes mellitus or renal failure should be differentiated at the outset. A primary inability to concentrate urine with secondary polydipsia is known as diabetes insipidus (DI), and may be due to either impaired release of AVP from the posterior pituitary (cranial diabetes insipidus, CDI) or to impaired renal response to AVP (nephrogenic diabetes insipidus, NDI).

The major causes of CDI and NDI are shown in Box 4.5. There are four causes of inherited CDI, all extremely rare. Autosomal dominant CDI is caused by mutations involving the pre-pro-vasopressin gene excluding the region coding for AVP itself. The disorder is due to

BOX 4.5 Causes of diabetes insipidus**Cranial diabetes insipidus***Hereditary*

- Autosomal dominant
- Autosomal recessive
- Wolfram (DIDMOAD) syndrome
- Congenital septo-optic dysplasia.

Acquired

- Traumatic
- Post-hypophysectomy
- Tumour (primary or secondary)
- Granuloma (e.g. sarcoid, tuberculosis, histiocytosis)
- Infection
- Vascular impairment
- Autoimmune
- Idiopathic

Nephrogenic diabetes insipidus*Hereditary*

- X-linked (loss of function mutation AVPR2)
- Autosomal (loss of function mutation AQP 2)

Acquired

- Hypokalaemia
- Hypercalcaemia
- Drug induced (e.g. lithium, demeclocycline, methoxyflurane, amphotericin)
- Renal disease (e.g. medullary cystic disease, obstructive uropathy)

Vasopressinase-related diabetes insipidus

- Pregnancy

a progressive degeneration of magnicellular neurons, probably due to the accumulation of abnormal AVP-neurophysin II complexes. Presentation may be in the neonatal period or delayed until later in childhood. An even rarer autosomal recessive form exists in which the coding for AVP is mutated; presentation is in the early neonatal period. The two other causes of inherited CDI

are Wolfram syndrome and congenital septo-optic dysplasia, with about one-third of patients with each condition exhibiting overt symptoms of CDI. Wolfram syndrome may present with a combination of diabetes insipidus, diabetes mellitus, optic atrophy and deafness – hence the alternative name of DIDMOAD syndrome. Patients with either CDI or NDI may retain a partial ability to concentrate the urine, which can lead to diagnostic confusion. The osmoregulatory set point for AVP release and thirst is maintained in both conditions. Thus, in developing CDI, as the number of active cells releasing AVP from the posterior pituitary diminishes, the slope of the osmoregulatory response will decrease (see Fig. 4.5). With a plasma osmolality for thirst normally set slightly above that for AVP release, the urine concentration that can be achieved at the onset of thirst will diminish. When CDI is severe, thirst will be stimulated even though the urine is close to maximum dilution. Intermediate stages can, therefore, provide confusion, as an increase in plasma osmolality due to water deprivation despite thirst can be associated with a normally concentrated urine. Diagnostic confusion may be further compounded by two additional features not shown in Figure 4.5:

- an upregulation of renal responsiveness to AVP in CDI in which the kidney becomes more responsive to circulating vasopressin
- a secondary loss of urine-concentrating ability once the sustained polyuria has resulted in medullary washout, with a reduction of the concentration gradient within the medulla of the kidneys.

The former phenomenon may mask a developing CDI. The latter feature may result in a very blunted response to endogenous and exogenously administered AVP (or analogue), thus suggesting a primarily nephrogenic basis to the polyuria, whereas in fact the loss of renal concentrating ability is secondary to the severe polyuria.

Congenital NDI is a rare condition. About 90% of cases are X-linked and due to a loss of function mutation within the gene coding for the V_2 AVP receptor (AVPR2). Affected males present within the first six months of life with failure to thrive, irritability, severe thirst and excessive wetting. Often the diagnosis is made during the

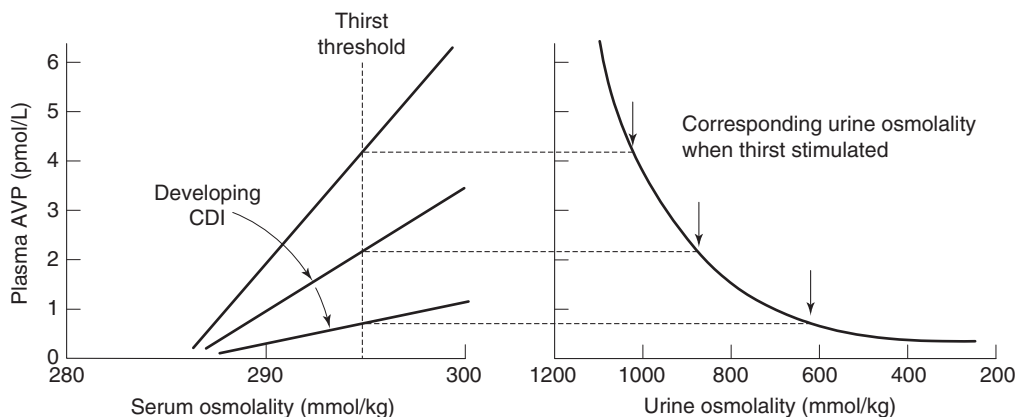


FIGURE 4.5 ■ The relationship between the osmoregulatory control over arginine vasopressin (AVP) secretion and the renal response to its action. For illustrative purposes, thirst is shown as a single point threshold. Normally, thirst does not become active until urine concentration approaches maximum. In developing cranial diabetes insipidus (CDI), as the number of cells releasing AVP diminishes, so thirst becomes active at correspondingly reduced urine osmolality. Polyuria leads to polydipsia.

first acute hospital admission because of impaired consciousness or convulsions, due to severe water depletion leading to significant hypernatraemia. The kidneys are completely unresponsive to AVP or AVP analogues. The female carriers of the condition are of interest, as many show evidence of polyuria, especially during pregnancy. Originally, this polyuria was considered to be due to the mimicking of affected males within the family, but this explanation is now disproven, with many females having demonstrably impaired renal responses to AVP – a partial form of congenital NDI – but some being almost as severely affected as males with congenital NDI. Of the remaining 10% of congenital NDI, inherited as autosomal recessive conditions, about 7% are due to mutations in the AQP 2 gene and the remaining 3% currently have no identified cause.

Both lithium and demeclocycline may produce clinically severe NDI with marked degrees of polyuria. Patients receiving lithium should be monitored strictly, to ensure effective but safe plasma concentrations of lithium to reduce the incidence of unwanted effects, including polyuria. Demeclocycline, once used to treat acne vulgaris, is no longer used as an antibiotic in man, but has, in the recent past, occasionally been used to control water-retaining states (see p. 51). The NDI of hypokalaemia and hypercalcaemia is generally less severe, with polyuria rarely greater than 4L/24h in adults. Recent work has suggested that NDI secondary to lithium, hypokalaemia and obstructive uropathy may all be related to alterations in AQP 2 expression.

Pregnancy and polyuria

There are several subtle alterations to normal osmoregulation in pregnancy, with the osmotic thresholds for both thirst and AVP secretion falling ~10 mmol/kg around the 6th week of gestation and remaining at these lower levels throughout pregnancy, so that a new steady state is achieved. The postpartum rise of these thresholds to normal occurs within approximately three weeks.

The major potential complication of osmoregulation in pregnancy is polyuria with secondary polydipsia. Pregnancy is discussed separately from other causes of polyuria because, although partial forms of CDI and NDI may be unmasked, one underlying cause is unique to pregnancy. This is due to the massive increase in the clearance of AVP, in both the placenta and the circulation, as a result of the release from the placenta of vasopressinase (cystine aminopeptidase, oxytocinase). In addition, there is evidence of increased clearance of AVP by the liver and kidneys related to increased blood flow to both organs. The increased clearance of AVP produces a form of diabetes insipidus that is unresponsive to exogenous pharmacological doses of AVP, but is responsive to the now commonly prescribed vasopressin analogue desamino-8-D-arginine vasopressin (dDAVP, desmopressin), which is resistant to enzymic degradation by vasopressinase. The response to a standard water deprivation test is thus identical to that found in CDI. Diabetes insipidus related to excessive vasopressinase activity resolves quickly following birth, with normal osmoregulation resuming within two weeks.

Polyuria secondary to primary polydipsia

Primary polydipsia is found in association with a variety of psychiatric and other CNS disorders. The commonest form is the ‘compulsive water drinking’ found in up to 7% of psychiatric inpatients. One possibility is that these patients have a reduced osmoreceptor setting for thirst, which is at or below the setting for AVP release. However, a subgroup of compulsive water drinkers use excessive water intake as a form of purging or self-therapy, and thus formal assessment of thirst thresholds using subjective visual analogue scales may be difficult, if not impossible, to interpret. This latter group has sometimes been classified as having psychogenic polydipsia, although the terms ‘compulsive water drinking’ and ‘psychogenic polydipsia’ are often used synonymously. The mechanisms are further complicated by the fact that many of the drugs used in psychiatry have powerful anticholinergic effects and may cause dryness of the mouth with a subsequent desire for oral fluids.

Primary polydipsia due to a hypothalamic disorder is rare, but can be dramatic. Even immediately following ingestion of large volumes of water, patients may possess such cravings for further water intake that they will go to extreme lengths to obtain it, such as consuming bath water or even lavatory water. Hypothalamic disorders can occasionally give rise to a water-retaining state and, if excessive thirst is also present, severe acute hyponatraemia may result (see p. 46).

Excessive thirst is a rare feature of conditions in which the renin–angiotensin system is stimulated, possibly because the concentration of angiotensin II, a powerful central thirst stimulant, is increased. Thus, pathological thirst has been described in renal artery stenosis, Wilms tumour, end-stage renal disease and even congestive cardiac failure. There is some evidence that when thirst occurs in these conditions, it can be controlled by angiotensin-converting enzyme (ACE) inhibitors. The mechanisms, however, are not well understood, as paradoxically, some animals develop increased thirst in response to ACE inhibitors; this is presumed to be because of increased angiotensin I delivery to thirst-stimulating areas of the brain that are inaccessible to ACE inhibitors, and at which conversion to angiotensin II subsequently occurs.

Laboratory investigation and treatment of polyuria

The first step is to confirm polyuria. Patients (and doctors) may confuse frequency of micturition with polyuria. True polyuria is invariably associated with polydipsia, and nocturnal polydipsia is an important pathological symptom. Fasting plasma glucose, serum potassium, calcium and urea measurements in the basal state may reveal diabetes mellitus, hypokalaemia, hypercalcaemia or renal impairment. Serum sodium may provide a pointer to the final diagnosis, tending towards the upper reference limit in secondary polydipsia and DI, but towards the lower reference limit in primary polydipsia. However, overlap of serum sodium concentrations between the diagnostic groups means that a water deprivation test is often necessary.

Water deprivation test. A standard protocol that can be performed on outpatients or day cases in the majority of patients is provided in [Appendix 4.2a](#). The test is conceptually simple to understand but is often incorrectly performed. Weighing the patient prior to and during water deprivation is a most important component of the test. Weighing allows the detection of surreptitious water intake – weight loss during the test should approximate to urine output (1 L = 1 kg). Weighing is also necessary to prevent gross dehydration in patients with severe DI; a weight loss of 3% body weight is normally the maximum that should be allowed, and if this is achieved, the test is concluded with dDAVP administration. However, care should be applied with respect to the weight loss limit as certain patients, often but not invariably with primary polydipsia, may prepare for the test by considerable water loading. Alternatively, patients with severe DI may start the test grossly dehydrated. Basal serum and urine osmolality should be measured before the test is started, as a low serum sodium and/or osmolality ($[Na]_s < 135$ mmol/L, $[Osm]_s < 275$ mmol/kg) may imply prior water loading and therefore the limit on weight loss can be extended, whereas an elevated serum sodium and/or osmolality ($[Na]_s > 145$ mmol/L, $[Osm]_s > 295$ mmol/kg) with dilute urine at the start of the test should preclude initiation of water deprivation.

If the water deprivation part of the test is concluded without adequate urine concentration being achieved (urine osmolality < 600 mmol/kg or urine:serum osmolality ratio $< 2:1$), dDAVP should be administered intravenously or intramuscularly. A sample of urine is collected after 2 h and, if necessary, at intervals during the next 16 h to determine maximum achievable urine concentration. Patients are allowed to drink during this time, but intake should be limited: a total input over 16 h of not more than 1000 mL is recommended. This is particularly important in patients who have primary polydipsia, as otherwise water intoxication could ensue. For patients with severe NDI, however, this limit may be inadequate and some extension may be required, matching output to input. Problematic patients should, therefore, be admitted to hospital and kept under strict observation.

Some care is required in interpretation of the final results of urine concentration achieved following dDAVP. Any patient with a history of prolonged polyuria of whatever cause will have some impairment of urine concentrating ability because of renal medullary washout. Attempts have been made to establish diagnostic criteria based upon the percentage increase of urine concentration following exogenous AVP preparations and that achieved by water deprivation alone. Such fixed criteria do not take into account serum osmolality measurements during or after water deprivation and may lead to misclassification. The majority of patients under investigation will be patients who have CDI or primary polydipsia. These patients may only achieve urine concentrations of 400 mmol/kg following dDAVP but, in the case of CDI, serum osmolality at the end of water deprivation would be at or above the physiological range. If diagnostic confusion exists, then further information may be achieved by plasma AVP measurements

during a repeat water deprivation test. Alternatively, a hypertonic saline infusion with serial plasma AVP measurements may be necessary.

Hypertonic saline infusion. A protocol for hypertonic saline infusion is given in [Appendix 4.2b](#). This test involves producing an acute elevation in serum sodium (~ 10 mmol/L) with frequent monitoring of plasma AVP concentration. Although acute extracellular osmotic changes should generally be avoided, the degree of change induced by the procedure is sufficient to induce significant and easily detectable increases in plasma AVP without complications; the morbidity of the test is extremely low. The major function of this test is to differentiate partial forms of CDI from primary polydipsia and from partial NDI. In CDI, the slope of the relationship between the plasma osmolality and the plasma AVP is reduced, but thresholds for thirst and AVP release are normal (see [Fig. 4.5](#)). The slope of this relationship remains normal in primary polydipsia and NDI.

Management of polyuria

The treatment of CDI has been revolutionized following the introduction of dDAVP. This analogue of AVP has a prolonged action and virtually no pressor activity. It is simple to administer either orally or nasally (by metered dose nasal spray). Patients tolerate the treatment extremely well and soon become practised in altering doses to meet their social requirements. Dilutional hyponatraemia is a potentially serious side-effect, but, in practice, is rarely encountered.

Because of the efficacy and patient acceptance of dDAVP, other forms of treatment have virtually disappeared. Lysine vasopressin (porcine antidiuretic hormone, available in oil as Pitressin) suffers the major disadvantage of native AVP in that, to achieve reasonable duration of action, a dose high enough to produce pressor effects may be required. Other drugs, which act by either enhancing renal responsiveness to AVP, for example chlorpropamide, or inhibiting renal diluting capacity, for example thiazide diuretics, are no longer used for CDI.

Primary polydipsia is a difficult condition to treat, but, in many instances, requires no primary treatment other than reasonable restriction of access to water. The major problem in this condition is the avoidance of antidiuresis and thus the potential for acute water intoxication.

The treatment of NDI is dependent on the cause. Acquired forms, such as those resulting from hypokalaemia or hypercalcaemia, require correction of the primary metabolic disturbance. Lithium treatment is probably the commonest cause of NDI, although the prevalence has fallen considerably since the recommended therapeutic range for serum lithium in the treatment of bipolar affective disorder has been reduced. Some reduction of polyuria can often be attained by reduction of lithium dose or by dividing doses of lithium to avoid very high peak serum concentrations. The renal resistance to AVP is related to serum lithium concentration. For intractable cases of lithium-induced polyuria, alternatives to lithium therapy should be sought.

Congenital NDI poses considerable problems of management. The severe polyuria may lead to enlarged bladder, hydronephrosis and hydronephrosis, and eventually to renal impairment, which ironically will reduce the symptoms. In addition, any failure to respond to the secondary polydipsia, such as may occur during illness or any degree of incapacity, will lead rapidly to hypernatraemia. The main treatment available until now has been thiazide diuretics, which inhibit the function of the diluting segment of the distal tubule, coupled with sodium restriction (<1.5 mmol/kg per 24h). This treatment can be used in combination with amiloride in older children, but is less well tolerated in the infant. Alternative combinations include a thiazide with a non-steroidal anti-inflammatory drug such as indometacin or a combination of thiazide with a cyclooxygenase-2 inhibitor. Treatments such as these can reduce urine flow rates to about one-third of pretreatment levels. New treatments are now under active investigation. Many of the loss of function mutations of AVPR2 lead to intracellular retention of the receptor (impaired trafficking to the basolateral membrane) rather than actual reduced affinity for AVP. Molecular chaperones, also known as 'pharmacochaperones', show promise as future therapies for such mutations. In addition, AVP-independent targeting of AQP 2 insertion into the apical membrane by selective E-prostanoid receptor agonists, may also prove to be a useful therapy.

Nocturnal polyuria

The study of renal diurnal rhythms of water and electrolyte excretion consistently demonstrates that, in health, both urine flow and electrolyte excretion are attenuated during sleep. Thus, it is expected that, unless significant water and/or electrolyte loading occurs immediately prior to retiring to bed, physiologically there will be no interruption to sleep due to the need to urinate. The terms nocturia and nocturnal polyuria are poorly defined. Nocturia as a descriptive symptom simply means the voluntary voiding of an indeterminate volume of urine at night (as opposed to enuresis, which is involuntary voiding). As a symptom, it can be extremely debilitating, as the frequent and persistent interruption of sleep can lead to severe fatigue. Nocturia is a common complaint of increasing prevalence with age in both men and women. One reason for the lack of consistency in the definition or investigation of nocturia, and in particular nocturnal polyuria, is that patients can present to a variety of medical and surgical specialties dependent on age, gender and on associated conditions, symptoms and subjective assessment.

Any disease process that reduces the functional capacity of the bladder or reduces the ability of the kidney to concentrate urine may give rise to daytime frequency and/or polyuria and nocturia. For some patients, however, nocturia is not associated with either daytime frequency or polyuria, but is associated with relative polyuria only at night. **Box 4.6** lists the common causes of nocturnal polyuria.

BOX 4.6 Causes of nocturnal polyuria

Sequestered fluid excreted at night

- Congestive heart failure
- Nephrotic syndrome
- Autonomic failure
- Severe varicose veins

Failure to conserve fluid at night

- Poor control of diabetes mellitus at night
- Suboptimal treatment of diabetes insipidus at night
- Non-dipping hypertension (sustained hypertension at night)
- Obstructive sleep apnoea
- Chronic renal failure

Laboratory investigation and treatment of nocturnal polyuria

To diagnose nocturnal polyuria, it is first necessary to exclude daytime polyuria and urodynamic causes of frequency. The investigation of the osmoregulation of a patient with suspected nocturnal polyuria is difficult, not least because there are no universally accepted investigation protocols. Moreover, it is likely now that investigations would be expected to take place in an outpatient setting, rather than on a fully staffed and equipped clinical investigation unit. The simplest demonstration of the existence and severity of nocturnal polyuria is to perform a 12 h-split 24 h urine collection, with one 12 h session incorporating the whole period of sleep, for example collection one 09.00–21.00 h, collection two 21.00–09.00 h. A 12 h urine collection overnight normally contains approximately one-half of the total 24 h creatinine excretion, but much less than half the volume of overall solute excretion. More detailed study is possible but requires a high degree of patient cooperation in the outpatient setting. Patients can be instructed to collect a complete 24 h urine collection, while maintaining an accurate diary of voiding times, collecting individual aliquots of each voiding and, if possible, an accurate volumetric measurement of each voiding. Alternatively, estimates of the volume of each voiding can be made from the relative creatinine concentration of each timed voiding, particularly if the patient is willing to reduce diurnal variations in creatinine excretion by avoiding meat consumption over the period of study. Providing the patient is willing and capable of following such a protocol, detailed diurnal patterns of excretion of volume, sodium and osmoles (or, indeed, any measurable urine constituent) can be constructed. **Figure 4.6** demonstrates the diurnal pattern of sodium excretion in a man presenting with severe nocturnal polyuria and a 12 h split urine volume of 0.57 L (daytime) to 1.75 L (night-time). This man was subsequently shown to suffer from severe obstructive sleep apnoea. Following treatment with continuous positive airways pressure (CPAP), the nocturnal polyuria resolved completely as the nocturnal natriuresis was reversed. Other causes of nocturnal polyuria, for which treatment of the primary cause is not possible or is insufficient to ameliorate symptoms, have been managed either with a loop diuretic taken 6 h prior to retiring to

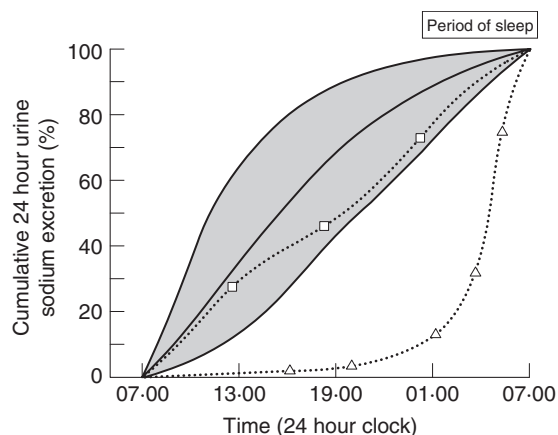


FIGURE 4.6 ■ Diurnal pattern of urine sodium excretion in a man presenting with severe nocturnal polyuria due to obstructive sleep apnoea. The shaded area shows the mean and 95% confidence limits of normal cumulative 24 h sodium excretion (...△..., pre-treatment; ...□..., post-treatment with continuous positive airways pressure).

bed (to induce a relative sodium depletion at night) or in combination with night-time dDAVP.

Hypernatraemia

Mild hypernatraemia (serum sodium >145 mmol/L) is not uncommon in hospital patients, especially in the elderly, but severe hypernatraemia (>154 mmol/L) has an incidence of less than 0.2% of hospital admissions per annum. The causes of hypernatraemia are shown in [Box 4.7](#).

Spurious hypernatraemia may occur due to sampling error, for example sampling from the same limb receiving a hypertonic sodium infusion such as 8.4% sodium bicarbonate. Spurious hypernatraemia may also occur because of post-collection contamination of the sample with sodium salts or because of evaporative losses after collection.

Acute sodium poisoning may also result in severe hypernatraemia (see [p. 38](#)), which is not initially due to water depletion, but will induce water depletion through

BOX 4.7 Causes of hypernatraemia

Spurious

- Sampling error
- Contamination
- Evaporation

Acute sodium poisoning

- Oral
- Parenteral

Water deficiency with unsatisfied thirst

- With polyuria (cranial diabetes insipidus, nephrogenic diabetes insipidus)
- Without polyuria (confusion, coma, immobility)

Water deficiency without conscious thirst

- Hypodipsic hypernatraemic syndromes

natriuresis and diuresis, and will require correction with free water administration.

The major clinical causes of hypernatraemia are due to water deficit, either physiologically, when the desire to drink cannot be expressed, e.g. in neonates, the old or incapacitated or unconscious patients, or pathologically, e.g. with a history of polydipsia and thirst that is, for whatever reason, not satisfied.

Water deficiency with thirst

Adult patients with water deficiency, whose ability to ingest water is impaired by incapacity, will be oliguric (<400 mL/24h). The urine will be concentrated (osmolality >1000 mmol/kg), although the capacity for urinary concentration is restricted in the neonate to ~ 500 mmol/kg and, in the adult, urine concentrating capacity declines in the elderly and may fall to <700 mmol/kg over the age of 70.

In patients whose water deficiency is due to complete DI (cranial or nephrogenic), the urine is not concentrated. However, if the water deficiency is sufficient to reduce ECF volume and thus renal function significantly, the urine may not be maximally dilute, but will approximate to the osmolality of serum. In addition, in partial forms of DI, severe water deficiency may result in near maximally concentrated urine. In order to avoid diagnostic confusion, it is, therefore, advisable to correct any water deficiency prior to establishing the diagnosis with a water deprivation test.

Water deficiency without thirst

Patients rarely may present with severe hypernatraemia but no thirst or polyuria. The cause is a group of conditions known collectively as adipsic hypernatraemia and hypodipsic or essential hypernatraemia. The hypothalamic disorder causing hypodipsic hypernatraemia may be due to trauma, a primary or secondary tumour, a granuloma (e.g. sarcoid or histiocytosis) or vascular impairment, or the condition may be idiopathic. The aetiological similarity to those conditions causing CDI underlines the close proximity of osmoreceptors for AVP and thirst control within the hypothalamus. Four subtypes of hypodipsic hypernatraemia have been recognized ([Fig. 4.7](#)).

In type 1 ([Fig. 4.7, line A](#)), normal osmotic control over AVP release is retained but thirst appreciation is absent – primary adipsia. This disorder is rare, but demonstrates the separate osmoregulation of thirst and AVP. An example of this defect in a male patient with a history of chronic hypernatraemia is shown in [Figure 4.8](#). During hypertonic saline infusion, he retained normal osmotic control over plasma AVP release, but subjective thirst assessment remained absent throughout. He had undergone hypothalamic surgery for severe behavioural disturbance and his serum sodium was found to be consistently >150 mmol/L without any experience of thirst. Adipsic hypernatraemia has been described in up to 20% of patients surgically treated for craniopharyngioma.

In type 2 hypodipsic hypernatraemia ([Fig. 4.7, line B](#)), the plasma AVP response is similar to that seen in patients with severe CDI, with a substantial loss of incremental gain of plasma AVP as serum osmolality rises.

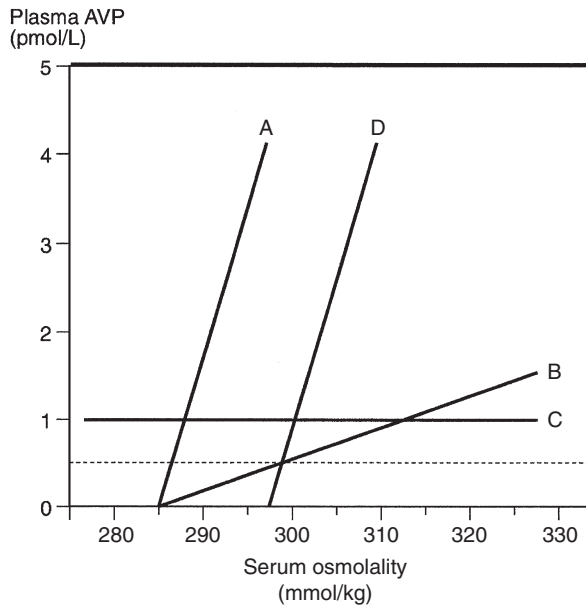


FIGURE 4.7 ■ Relationship between plasma arginine vasopressin (AVP) and serum osmolality in patients with hypodipsic hypernatraemia: (A) normal regulation of AVP release; (B) partial defect in AVP release; (C) total defect; (D) reset osmostat. The dashed line represents the limit of detection of current AVP assays. Adapted from Robertson et al. 1982.

Unlike the situation in CDI, however, the osmotic thirst response is severely blunted. This group of patients is of interest, as the renal responsiveness to AVP is often considerably enhanced and patients are thus not polyuric. Urine concentration may approach the maximum when hypernatraemia is severe and patients retain the ability to dilute urine.

In type 3 (Fig. 4.7, line C), plasma AVP concentration is unresponsive to changes in serum osmolality, but is fixed at a low level. Osmotic thirst is entirely absent. Patients are, therefore, at risk of developing either

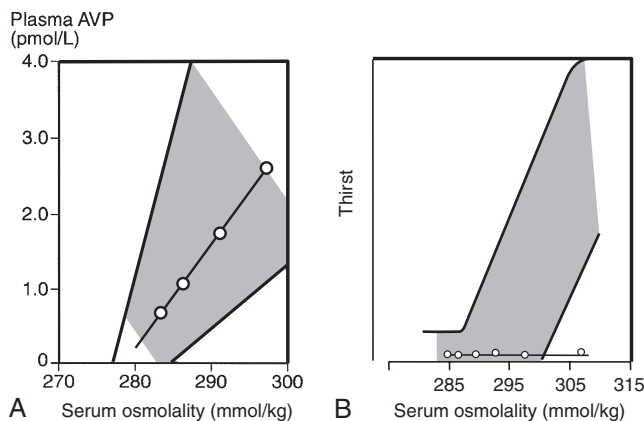


FIGURE 4.8 ■ Osmoregulation in a patient with type 1 hypodipsic hypernatraemia syndrome. (A) Normal osmotic control of arginine vasopressin (AVP) release. (B) Complete absence of thirst stimulation despite significant osmotic stimulus. Shaded areas represent 95% confidence limits of normal response.

hypernatraemia or hyponatraemia, depending on the prevailing fluid intake.

Type 4 hypodipsic hypernatraemia (Fig. 4.7, line D) represents true resetting of the osmostat for both plasma AVP response and thirst: the incremental gain for both remains normal and patients can concentrate and dilute urine normally, but around a higher setting of plasma sodium and hence serum osmolality. Severe hypernatraemia is generally not a feature, as dehydration does result in thirst. Thus, hypodipsia in these patients is only relative to the degree of thirst seen in normal subjects with corresponding degrees of hypernatraemia; the full range of thirst appreciation is maintained.

Management of hypernatraemia

Hypernatraemia in the incapacitated patient is invariably due to water depletion and should be corrected. Care should be exercised in the rate of correcting severe hypernatraemia in which the duration is greater than 48 h, and the rules on the rate of correction are directly comparable with those applicable to the correction of chronic dilutional hyponatraemia (see p. 51). The response of brain cells to persistent severe dehydration is to generate intracellular solutes, sometimes termed idiogenic osmoles or osmolytes. These are organic solutes that can be synthesized without apparent detrimental effect on cell structure or function. Recent nuclear magnetic resonance spectra of hypernatraemic human brains in vivo have identified myoinositol, scylloinositol, N-acetylaspartate and choline as osmolytes. Oral hydration with plain water is the simplest regimen (for estimate of deficiency, see Appendix 4.1c), but if oral or enteral fluids cannot be given, then hypotonic sodium or isotonic glucose solutions should be infused. It is strongly recommended that serum sodium is corrected at a rate of no greater than 12 mmol/L per 24 h; once corrected, any underlying polyuria should then be investigated as outlined in the previous section. A conscious patient with hypernatraemia, but without expressed thirst, is by definition suffering from one or other form of hypodipsic hypernatraemic syndrome. Although it is possible to differentiate pure adipsia (type 1) from other types of the syndrome by the use of tests of water deprivation and water loading (see Appendix 4.2), in practice the diagnosis can only be confirmed by the study of plasma AVP secretion and thirst in response to hypertonic saline infusion.

Management of hypodipsic hypernatraemia syndromes. The treatment of hypodipsic hypernatraemic syndromes varies considerably with type. Type 1, in which AVP release is normal but thirst is absent, and type 2, with a reduced pituitary release of AVP, enhanced renal response to AVP and blunted thirst response, both require a fluid intake regimen of more than 2 L/24 h for an adult to avoid hypernatraemia. Concentration of the urine should be avoided and at least one dilute urine should be passed each day. Type 3 poses the greatest problem of management, as fluid output and urine concentration cannot be used by the patient to gauge fluid balance. A standard regimented fluid intake is required and body weight needs to be accurately monitored each day to

monitor fluid status. Frequent plasma monitoring is required to avoid the progressive development of hyper- or hyponatraemia. Type 4 requires no specific treatment, as alterations in fluid regimens will simply be handled in a physiologically normal way except at a higher osmotic setting than in normal individuals.

Hyponatraemia

Hyponatraemia (serum sodium concentration <130 mmol/L) is common in hospital inpatients, with a prevalence of 2–3%. The lower reference limit for serum sodium is reduced by ~ 5 mmol/L in hospitalized patients as opposed to healthy controls. The majority of cases of hyponatraemia are mild, self-limiting and require neither treatment nor extensive investigation.

By definition, hyponatraemia implies an excess of water in comparison with sodium, although primary water excess is not always responsible for hyponatraemia. For example hyperglycaemia or the infusion of mannitol may induce significant fluid shifts from the ICF to the ECF. An estimate of the expected fall in serum sodium in relation to hyperglycaemia is given in [Appendix 4.1d](#). However, this section is primarily concerned with states of water excess and with acute and chronic dilutional hyponatraemia: dilutional hyponatraemia is also defined as hyponatraemia with clinically normal ECF volume or normovolaemic hyponatraemia. The full classification of hyponatraemia is shown in [Box 4.8](#).

Spurious hyponatraemia is due to *in vivo* or *in vitro* contamination of the specimen with water or fluid containing sodium at a concentration that is less than that of plasma. The commonest example seen in hospital practice is venesection from a limb concurrently receiving a dextrose infusion.

Pseudohyponatraemia (molar concentration) is due to the replacement of a portion of the plasma water space with either lipid or protein. This can occur in patients with severe endogenous or exogenous hypertriglyceridaemia and in patients with high plasma protein concentrations (usually due to paraproteinaemia). The concentration of sodium in plasma water (molal concentration) is normal, but each litre of plasma now contains less water. Serum osmolality (molal concentration) is normal.

Patients with hypovolaemic or hypervolaemic hyponatraemia present with signs of ECF volume depletion

or excess. These conditions are discussed under disorders of sodium metabolism; they are usually differentiated clinically from dilutional hyponatraemias and also by the presence of biochemical evidence of the underlying cause, for example renal or hepatic dysfunction.

Although acute and chronic hyponatraemia may share certain common causes and the definition is somewhat arbitrary (acute hyponatraemia is of <48 h duration), these conditions present and are managed in quite separate ways. This is because of the adaptation of cells, in particular brain cells, to chronic hypotonicity. The nature of this adaptation is summarized in [Figure 4.9](#) and can, in part, be attributed to the reduction of cerebral osmolytes (see [p. 45](#)).

Acute dilutional hyponatraemia

The main causes of acute dilutional hyponatraemia are shown in [Box 4.9](#). Self-inflicted acute hyponatraemia has been described, but is extremely difficult to induce due to the enormous human capacity for renal water excretion – urine flow rates of 27.5 mL/min (equivalent to 39.6 L of urine in 24 h) have been described, although low osmotic loads for excretion or renal impairment can severely reduce water clearing capacity.

Psychogenic polydipsia (and the much rarer hypothalamic polydipsia) may result in acute hyponatraemia. However, the commonest cause of acute dilutional hyponatraemia seen in hospitalized patients is inappropriate postoperative fluid regimens involving the intravenous infusion of excessive volumes of low-sodium fluid (e.g. ‘dextrose saline’). The acute hyponatraemia following transurethral prostatectomy (TURP), known as the transurethral resection syndrome (TURS), is also worthy of specific mention, as it is the most serious acute hyponatraemia likely to present at the current time. In the development of TURS, irrigant fluid (1.5% glycine) is absorbed into the circulation via the open venous sinuses of the prostatic bed; glycine is used as it provides a non-electrolytic medium that will not dissipate the energy of cautery used in the resection, and the slightly hypotonic concentration provides an ideal optical interface. Risk factors for the development of the syndrome include a prolonged operative procedure, high irrigant hydrostatic pressure and significant blood loss. Strict control over these risks has resulted in a marked reduction in the incidence of this complication in recent years. If, however, sufficient irrigant is absorbed, this fluid will reduce plasma sodium concentration by simple dilution, but osmolality will not initially be reduced by an equivalent extent. An osmolal gap (see [Appendix 4.1e,f](#)) will initially develop as a result of high plasma glycine concentrations (up to 15 mmol/L), but unless the resultant free water is subsequently excreted, the osmolal gap will disappear as the glycine is metabolized. High plasma concentrations of both glycine and ammonia (which may result from the rapid metabolism of glycine) have been proposed to contribute to the severe deterioration in mental state that occurs in full-blown TURS. However, patients with hypoosmolal hyponatraemia respond rapidly to measures designed to treat acute water intoxication alone.

BOX 4.8 Classification of hyponatraemia

- Spurious (sampling error)
- Pseudohyponatraemia
- Hypovolaemic hyponatraemia
 - Renal sodium loss (diuretic excess, adrenal failure, sodium losing nephropathy)
 - Non-renal sodium loss (gastrointestinal, haemorrhage, burns)
- Hypervolaemic hyponatraemia
 - With oedema (cirrhosis, CCF, nephrotic syndrome)
 - Without oedema (acute or chronic renal failure)
- Normovolaemic hyponatraemia
 - Acute dilutional hyponatraemia (see [Box 4.9](#))
 - Chronic dilutional hyponatraemia (see [Box 4.10](#))

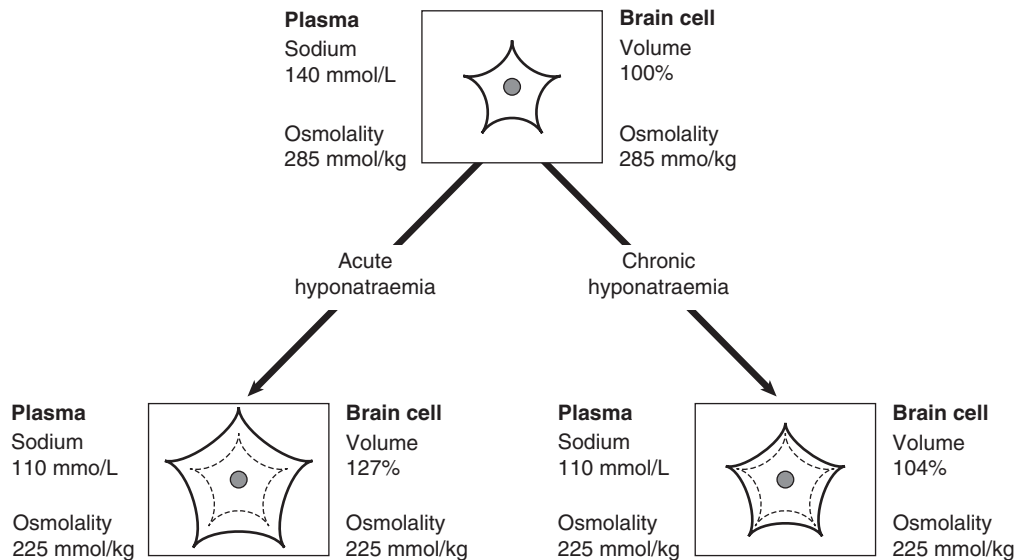


FIGURE 4.9 ■ The major differences in brain cell volume in acute hyponatraemia and chronic hyponatraemia. In chronic hyponatraemia, cell osmotic content has been reduced.

BOX 4.9 Causes of acute dilutional hyponatraemia

- Self-induced
- Psychogenic
- Freshwater drowning
- Iatrogenic
 - Inappropriate i.v. fluid regimen
 - Absorption of irrigant (transurethral prostatectomy, transcervical endometriectomy, transcutaneous ureterolithotomy, vesical ultrasonic lithotripsy)
- Syntocinon induction of labour with isotonic dextrose infusion

Acute hyponatraemia of sufficient severity to present clinically, with symptoms of headache, confusion and drowsiness, is a medical emergency, as the condition may lead to coma and death. Progressive clinical symptoms may be associated with signs of raised intracranial pressure, decerebrate posturing, fixed dilated pupils, bradycardia, hypertension and convulsions. All cells swell in size when placed in hypotonic medium: the change in size is particularly important with respect to brain cells confined within the fixed volume of the cranium (see Fig. 4.9). In acute hyponatraemia, adaptation of brain cell size cannot occur quickly enough to prevent the effects of cellular swelling and subsequent tentorial herniation with dramatic clinical sequelae. Females of reproductive age appear to be particularly prone to the effects of acute hyponatraemia.

Chronic dilutional hyponatraemia

Chronic dilutional hyponatraemia is common and has many possible causes (Box 4.10). The majority (excluding chronic hyponatraemia associated with beer potomania; see p. 49) can be related to a failure of water excretion,

presumed to be due to a failure to suppress AVP secretion. There is good evidence that certain tumours (e.g. small cell (oat cell) carcinoma of the bronchus) possess the capacity to synthesize and secrete AVP, but in the majority of cases, the origin of the AVP is presumed to be the hypothalamic-posterior pituitary axis.

Patterns of plasma AVP response to increasing plasma sodium concentration induced by hypertonic saline infusion have been studied in chronic hyponatraemia. Four patterns of response have been described, as shown in Figure 4.10.

1. In type A, there is no relationship between plasma AVP concentration and plasma sodium (osmolality); AVP is released either at fixed concentration or shows random variation. This pattern is often seen in association with ectopic production of AVP by tumours.
2. In type B, there is a normal response of AVP release to osmotic stimulation but it occurs at a much lower setting – a resetting of the osmostat. Theoretically, this condition could result from the adaptive response to chronic hyponatraemia as the osmotic content of brain cells falls. This condition has been described in association with tumours and various neurological disorders.
3. In type C, there is a constant low-level release of AVP, even when osmolality is suppressed below the normal threshold, but a qualitatively and quantitatively normal response above the threshold. This pattern may be seen transiently following pituitary/hypothalamic trauma.
4. In type D, there is suitable suppression of AVP concentration when the patient is hyposmolal, suggesting either that immunologically distinct antidiuretic material is being secreted or that the renal tubules are somehow rendered more sensitive to extremely low concentrations of circulating AVP. This pattern of response is seen in patients with gain of function mutations of the AVPR2 gene.

BOX 4.10 Conditions associated with chronic dilutional hyponatraemia**Tumours***Thoracic*

- Bronchogenic
- Carcinoma
- Bronchial carcinoid
- Mesothelioma
- Thymoma

Non-thoracic

- Pancreatic carcinoma
- Duodenal carcinoma
- Ureteral, bladder and prostatic carcinoma
- Lymphoma
- Myeloproliferative diseases
- Olfactory neuroblastoma

Pulmonary disease*Infections*

- Tuberculosis
- Empyema
- Lung abscess
- Pneumonias (bacterial, viral and mycoplasmal)
- Aspergillosis

Ventilatory

- Asthma
- Bronchiolitis
- Pneumothorax
- Emphysema
- Cystic fibrosis
- Positive-pressure ventilation

CNS disorders*Space-occupying lesions*

- Tumour (any type)
- Abscess
- Subdural haematoma

Degenerative

- Guillain–Barré
- Multiple sclerosis
- Cerebral atrophy

Infections

- Meningitis
- Encephalitis

Inflammatory conditions

- Systemic lupus erythematosus

Miscellaneous

- Head trauma
- Hydrocephalus
- Subarachnoid haemorrhage
- Cavernous sinus thrombosis

Drugs*AVP analogues*

- dDAVP
- Oxytocin/syntocinon

AVP potentiation

- Prostaglandin synthetase inhibitors, e.g. aspirin, indometacin
- Chlorpropamide

Stimulation of AVP release

- Nicotine
- Phenothiazines
- Vincristine
- Tricyclic antidepressants

Mixed or unknown action

- Colchicine
- Carbamazepine
- Clofibrate
- Cyclophosphamide
- Clozapine

Miscellaneous

- Idiopathic
- Gain of function mutation AVPR2
- Acute porphyrias
- Hypothyroidism
- Low osmotic load

Although this classification of AVP responses in chronic dilutional hyponatraemia assists pathophysiological understanding, no reliable relationship has been found between the aetiology of the condition and any one pattern of response, save for the gain of function mutations in the AVPR2 gene. The measurement of AVP in chronic dilutional hyponatraemia, therefore, for the vast majority of patients has no current diagnostic or prognostic function.

The syndrome of inappropriate antidiuretic hormone secretion. The term ‘syndrome of inappropriate antidiuretic hormone secretion’ (SIADH) was first coined by Bartter and Schwartz in 1957, to describe patients with severe hyponatraemia who were without evidence of renal failure, adrenal failure or saline depletion, but who had indirect evidence of persistent AVP secretion. The criteria for diagnosing the syndrome are shown

in **Box 4.11**. Before this description, such patients were often misclassified as suffering from renal salt wasting because of the associated hyponatraemia and natriuresis. The original description thus provided a useful explanation of the biochemical findings and led to more rational therapeutic regimens.

Unfortunately, the term SIADH tends to be used universally to describe any acute or chronic dilutional hyponatraemia. Apart from certain causes of acute water intoxication and rare forms of chronic hyponatraemia, plasma AVP is detectable in the vast majority of patients with hyponatraemia of whatever cause. In addition, the criteria for diagnosis may produce anomalies, for example chronic dilutional hyponatraemia with reset osmostat (**Fig. 4.10B**) could fit the criteria when plasma osmolality is above the new threshold for AVP release, but fail to comply when plasma osmolality falls below this threshold, when maximally dilute urine is excreted. The term

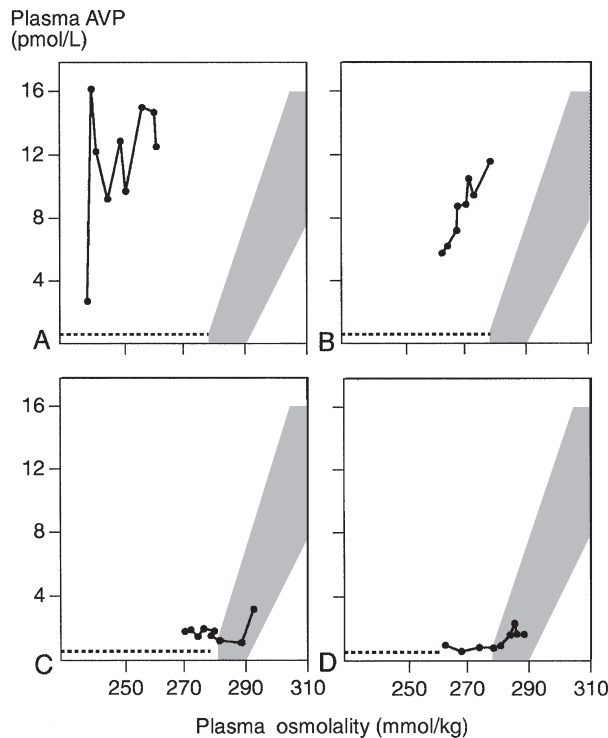


FIGURE 4.10 ■ The four patterns of plasma arginine vasopressin (AVP) responses to changes in plasma osmolality induced by hypertonic saline infusion in patients with chronic dilutional hyponatraemia. (A) Random variation. (B) Resetting of osmostat. (C) Constant low-level release. (D) Hypovasopressinaemic antidiuresis. Shaded areas represent the inter-subject variation limits of normal response. The dotted line represents the limit of detection of current AVP assays. Data from Zerbe R L, Stropes L, Robertson G L. Vasopressin function in the syndrome of inappropriate antidiuresis. *Annual Review of Medicine* 1980; 31:315–327.

BOX 4.11 Criteria for diagnosing the syndrome of inappropriate antidiuretic hormone secretion (SIADH)

1. Hyponatraemia with corresponding hypoosmolality of the serum and ECF
2. Continued renal excretion of sodium
3. Absence of clinical evidence of fluid volume depletion or overload, that is, normal skin turgor and blood pressure, or absence of oedema
4. Osmolality of the urine greater than that appropriate for the concomitant tonicity (effective osmolality) of the plasma, that is, urine not maximally dilute
5. Normal renal function
6. Normal adrenal function
7. Normal thyroid function

SIADH implies physiological understanding when often very little exists, and is a diagnosis that does not indicate immediate management or prognosis. Syndrome of inappropriate antidiuretic hormone secretion is a term that should now be replaced with more descriptive terminology, for example chronic dilutional hyponatraemia secondary to small cell carcinoma of the lung.

Sick cell syndrome. The association of sick cells and hyponatraemia was first explored by Flear and Singh in 1973, and it was an attempt to explore critically the assumption of common pathophysiology of the vast number of conditions grouped under SIADH, as well as to explain the hyponatraemia of seriously ill patients with conditions such as congestive cardiac failure (CCF), cirrhosis or serious pulmonary or CNS infections. One hypothesis was that sick cells leak normally non-diffusible solutes, but gain sodium, leading to extracellular hyponatraemia, but maintenance of serum osmolality. This combination of events may be observed in extremely ill patients after major surgery or burns, but has not been detected in the vast majority of patients with chronic hyponatraemia. However, a type of sick cell syndrome that is relevant to the development of chronic hyponatraemia is that resulting from a primary overall reduction in cellular osmotic content, through either increased loss or reduced production. Osmoreceptor cells so affected would result in control of AVP around a lower osmotic setting – a reset osmostat (SIADH type B). In practice, however, it is difficult to distinguish between a primary loss of cellular osmotic content (sick cells) or a secondary adaptive loss in response to a primary chronic dilutional hyponatraemia. Like SIADH, the term ‘sick cell syndrome’ should now be replaced by descriptive terminology, for example ‘chronic dilutional hyponatraemia secondary to chest infection’.

Low osmotic load hyponatraemia

One unusual cause of hyponatraemia requiring special mention is that associated with a low osmotic load for renal excretion. Beer potomania is an example of such a condition. It is caused by the consumption of a large volume of beer containing only very low quantities of electrolytes, coupled with a diet poor in protein and other minerals. Unlike starvation, in which osmoles such as ketones and urea are generated and excreted in the urine, the daily osmolal load for excretion in individuals at risk for beer potomania can be extremely low (<250 mmol/24h). Thus, if the minimum achievable urine osmolality is normal at about 50 mmol/kg, then, allowing for insensible fluid loss, a consumption of more than 6L of such beer over the 24h period will result in water retention independent of AVP secretion or responsiveness. Binges can result in acute hyponatraemia, although the more usual presentation is of a severe chronic hyponatraemia, occasionally complicated by plasma electrolyte changes secondary to alcoholic liver disease. The condition is unusual in the UK, probably because of the higher mineral content of traditional beers compared with continental beers, together with the tradition of serving bar snacks rich in sodium chloride. An association with excess cider consumption (cider potomania) has also been described, although this condition can only occur in an individual with normal urine diluting capacity if factory-type cider is consumed, rather than traditional craft cider. The former may have only the equivalent of 25% apple juice content, whereas the latter has 100% with a corresponding mineral content providing non-metabolizable osmoles in excess of 70 mmol/kg.

The concept of beer potomania as a cause of dilutional hyponatraemia is that of a mismatch between fluid intake and the capacity for excretion, independently of AVP secretion or responsiveness. This concept may have other parallels, for example in elderly patients treated with thiazide diuretics and with diets of low mineral content, adequate calories and relatively high fluid loads – the so-called ‘tea and toast’ diet. Thiazide diuretics act by blocking the reabsorption of sodium and chloride in the cortical diluting segment of the nephron. Thus, the capacity to generate maximally dilute urine is impaired and a concomitant ‘tea and toast’ diet can result in the development of severe dilutional hyponatraemia.

Cerebral salt wasting

The association of cerebral pathology, hyponatraemia and failure to retain urine sodium is termed cerebral salt wasting (CSW). This term was coined before the description and understanding of SIADH, in which there is also hyponatraemia and, during access to free water or following infusion of saline, a continued renal loss of sodium. Recent reports have attempted to re-implicate a primary renal sodium loss in some patients with cerebral pathology. Natriuretic peptide excess has been suggested as a mechanism. Caution is required, however, as many of such case reports invariably describe a euvolaemic, as opposed to hypovolaemic, hyponatraemia indistinguishable from hyponatraemia associated with SIADH. Differentiation of CSW from SIADH will become increasingly important as selective AVPR2 antagonists (vaptans) become more widely available and affordable (see p. 51). Vaptans are contraindicated in hypovolaemic hyponatraemia.

Laboratory investigation of hyponatraemia

Hyponatraemia is, by definition, a laboratory diagnosis, but the differentiation of hypovolaemic and hypervolaemic hyponatraemia is largely clinical, as is the differentiation between acute and chronic dilutional hyponatraemia.

Acute dilutional hyponatraemia may constitute a medical emergency, yet its laboratory investigation is often cursory and incomplete. The diagnosis is by clinical history of acute water loading without evidence of corresponding diuresis, and rapid deterioration in mental state, possibly coupled with signs of cerebral oedema. It is most important to confirm the serum sodium measurement, preferably with a fresh sample, ensuring that contamination of the specimen is avoided. The most important additional laboratory investigation is the measurement of serum osmolality, which will validate the serum sodium measurement or indicate the extent of any osmolal gap. Measurement of urine osmolality is superfluous, even if a sample is available.

The laboratory investigation of chronic dilutional hyponatraemia is predominantly to confirm the water-retaining state that has produced the condition and to exclude correctable metabolic causes. The serum electrolyte pattern in chronic dilutional hyponatraemia will usually show normokalaemia, but a hypochloraemia of corresponding degree to the hyponatraemia. The serum

urea concentration is often low and is reduced not only by dilution, but also because of reduced tubular reabsorption. Serum urate concentration may also be low because of changes in tubular function, but this is not invariable and renal handling of urate may be directly affected by the cause of the hyponatraemia, for example a thiazide diuretic.

Urine measurements may be of confirmatory help in chronic dilutional hyponatraemia, but can be misleading. The urine is inappropriately concentrated, that is, it is anything other than maximally dilute. There is widespread belief that urine osmolality must exceed serum osmolality for the diagnosis of a water retaining state, but this is not the case. The concentration of urine is dependent on both water and osmolal content. An anorexic patient may have a total osmolal load of only 400 mmol/24h to excrete, but if fluid intake is sufficient to produce 2 L of urine, then the overall urine concentration at balance is 200 mmol/kg. A urine concentration of 220 mmol/kg would produce a positive water balance, even though the urine is hypoosmolal to plasma.

One of the remarkable features of chronic dilutional hyponatraemia, especially when it is clearly demonstrated that plasma AVP concentrations are grossly elevated, is that patients do not form maximally concentrated urine. Patients with chronic dilutional hyponatraemia reach new steady states and then remain in sodium and water balance. Thirst thresholds appear to be downregulated to match the reduced plasma osmolality. Sudden increases or decreases in fluid intake may upset the steady state until a new one is reached. The mechanism of renal adaptation is probably related to downregulation of the expression of AQP 2 in the renal collecting ducts coupled to a reduction in the concentration gradient within the renal medulla.

The urine sodium concentration may also be of diagnostic value to differentiate hypovolaemic hyponatraemia if clinical signs are absent or misleading, but only if the sodium deficit is not renal in origin. Again, some care is needed, as chronic dilutional hyponatraemia is classically associated with a natriuresis, which can be defined essentially as anything other than maximal sodium retention. Often, when the patient is on free fluids, the natriuresis is very marked, with a urine sodium concentration >100 mmol/L. But as a new steady state is approached, urine sodium excretion falls, and when fluid and sodium intake are restricted, urine sodium may fall to very low concentrations (<10 mmol/L); this then indicates the total body sodium deficit that has developed secondarily to the natriuresis induced by initial positive water balance.

Other laboratory investigations recommended when the cause of the water-retaining state is not clear include thyroid function tests and a short tetracosactide stimulation test. There is a recognized, but variable, association of chronic hyponatraemia with hypothyroidism, and occasionally patients with primary or secondary adrenal failure present with nothing more specific than chronic dilutional hyponatraemia.

As previously stated, measurements of plasma AVP concentration are not normally of any assistance in the diagnosis of chronic dilutional hyponatraemia, but may

be of value to determine if normal osmoregulation has been restored following a period of treatment. An indirect assessment of water handling using a water load test can also occasionally be used for this purpose (see [Appendix 4.2c](#)).

Management of hyponatraemia

The management of hypovolaemic hyponatraemia is primarily to restore blood volume and ECF volume to normal and to correct the underlying disorder. The management of hypervolaemic hyponatraemia is primarily to treat the underlying disorder and to use diuretic therapy when appropriate. Management of both these conditions is covered within the sections on sodium deficiency and excess.

The management of euvolaemic or dilutional hyponatraemia has created considerable controversy in recent years. This has its origins in the morbidity and mortality of uncorrected severe dilutional hyponatraemia (serum sodium ≤ 115 mmol/L), whether it is acute or chronic in nature. In the USA, a regimen of partial correction of hyponatraemia was adopted in the past, which resulted in patients' serum sodium concentrations being increased rapidly to ~ 125 mmol/L. The rate of increase in sodium was usually in excess of 0.5 mmol/L per hour and often several-fold greater. Unfortunately, this regimen has become increasingly associated with the development of the neurological disorder known variously as central pontine myelinolysis (CPM), pontine and extrapontine myelinolysis or osmotic demyelination syndrome. The usual course of events in such cases is that the patient presents clinically with hyponatraemic symptoms, which can be either mild, such as weakness and confusion or severe, such as convulsions or coma. During the rapid correction of the plasma sodium concentration, these symptoms improve, but in the following few days the patient's neurological condition deteriorates and further symptoms develop, including behavioural disturbances and convulsions and these, in turn, may lead on to the full-blown condition of CPM, including pseudobulbar palsy and quadriparesis. Further confusion as to the origin of the condition has been a continued, but unsubstantiated, association in the literature of CPM with uncorrected hyponatraemia.

Recommended management of symptomatic acute dilutional hyponatraemia. The biggest danger in this condition is failure to treat promptly. Water restriction has a supportive role, but must not be used alone. Cerebral oedema is the major complication, and acute elevation of plasma sodium using 5% sodium chloride infusion to produce a *maximum* increase of 12 mmol/L per 24h or a serum sodium concentration of approximately 125 mmol/L (whichever is the lesser), is recommended. If an osmolal gap exists, then the infusion should increase serum osmolality to no greater than 255 mmol/kg. A formula to calculate the amount of sodium required is given in [Appendix 4.1g](#). Short-term rates of infusion resulting in an increase of serum sodium up to 5 mmol/L per hour, have, in acutely symptomatic patients, been used successfully

with dramatic clinical improvement and with no long-term sequelae. A simple general rule for patients falling within the definition of acute hyponatraemia and who are symptomatic is to match the rate of correction with the presumed rate of onset. Diuretics such as furosemide and mannitol may be of benefit, but lack the precision of control afforded by hypertonic saline infusion, and when used in combination with hypertonic saline infusion may make controlling the rate of correction more difficult.

Recommended management of chronic dilutional hyponatraemia. The cornerstones of management in chronic dilutional hyponatraemia are to prevent further falls in plasma sodium concentration, to treat any underlying condition, to alleviate any symptoms attributable to hyponatraemia and to avoid any therapeutic complications. Many patients who are in a steady state with mild hyponatraemia require no specific treatment. For those at risk of further falls in plasma sodium, water should be restricted to a degree sufficient to induce negative water balance (usually 500–800 mL of water per 24h). Water restriction will increase plasma sodium concentration in anyone, but the rate of increase may be very slow, and if any degree of sodium depletion exists, an increase in plasma sodium will only be achieved at the expense of a contraction in ECF volume and a potential deterioration in renal function. Certain patients cannot tolerate fluid restriction, and for these, until recently, treatment with the tetracycline antibiotic demeclocycline (DMC) may have been considered. Arginine vasopressin receptor antagonists (collectively termed 'vaptans') are now commercially available. Tolvaptan is a selective vasopressin V_2 -receptor antagonist with an affinity for the V_2 -receptor greater than that of native arginine vasopressin. Treatment of chronic dilutional hyponatraemia commences with a single 15 mg oral dose and can be titrated up to a single daily dose of 60 mg, depending on response. Regular monitoring of serum electrolytes and renal function is strongly recommended. Tolvaptan should not be used in the treatment of acute dilutional hyponatraemia, which requires hypertonic saline infusion. The introduction of selective AVPR2 antagonists has rendered previous treatment options of DMC or lithium, obsolete.

In the very rare instances in which severe chronic hyponatraemia produces neurological symptoms, it may be considered prudent to raise the plasma sodium concentration acutely with hypertonic saline. If this is considered, the rate of infusion should be most carefully controlled with frequent monitoring of serum sodium. As in the management of acute hyponatraemia, it is recommended that serum sodium is raised to a concentration no greater than 125 mmol/L by this method. In this instance, however, the rate of infusion should produce a change in serum sodium of no greater than 0.5 mmol/L per hour to minimize the risk of the development of CPM: this is in contrast to the treatment of acute symptomatic hyponatraemia, in which the recommended 24h rate and maximum concentration are the same, but the maximum rate for short periods may be much more rapid – up to 5 mmol/L per hour.

DISORDERS OF POTASSIUM METABOLISM

Potassium is predominantly an intracellular cation, but disorders of potassium metabolism are generally identified by measurements of extracellular potassium. Dramatic changes in plasma potassium concentration may occur as a result of transcellular shifts without necessarily any alteration of total body potassium. Clinical disorders are therefore classified on the basis of serum potassium concentration, rather than of potassium depletion or excess.

Hypokalaemia

The definition of hypokalaemia is not strict, but persistent serum potassium concentrations <3.4 mmol/L (plasma 3.1 mmol/L) require investigation. The immediate clinical effects of hypokalaemia are on neuronal and muscular function: they result from an increase in the ratio of the intracellular to extracellular potassium concentrations. In addition, hypokalaemia, when associated with severe potassium depletion, affects the function of a wide variety of organ systems. [Box 4.12](#) lists the possible clinical effects of hypokalaemia.

BOX 4.12 The clinical effects of hypokalaemia

Cardiovascular

- Electrocardiographic changes – depressed ST segment, inverted T waves, prominent U waves
- Arrhythmias
- Potentiation of digitalis intoxication
- Hypotension
- Myocardial necrosis

Neuromuscular

- Weakness, flaccid paralysis
- Pain, tenderness, cramps
- Tetany
- Rhabdomyolysis

Neuropsychiatric

- Lethargy, apathy, depression
- Acute memory loss, disorientation, confusion

Renal

- Polyuria
- Sodium retention
- Increased renal ammoniogenesis
- Metabolic alkalosis

Gastrointestinal

- Decreased motility (constipation, paralytic ileus)
- Decreased gastric acid secretion

Endocrine

- Increased renin secretion
- Decreased aldosterone secretion
- Decreased insulin secretion

BOX 4.13 Causes of redistribution hypokalaemia

In vitro redistribution

- Uptake by white blood cells (e.g. in leukaemia)
- Uptake by erythrocytes following in vivo insulin administration

In vivo redistribution

- Alkalosis
- Increased plasma bicarbonate
- Insulin administration
- β -adrenergic agonists
- Toxic chemicals (toluene, soluble barium salts)
- Hypokalaemic periodic paralysis

Causes of hypokalaemia

Hypokalaemia can be caused by redistribution of potassium (in vitro or in vivo), inadequate intake or excessive loss (renal or extrarenal).

Redistribution hypokalaemia in vitro. This form of hypokalaemia, sometimes known as spurious hypokalaemia, has been described in association with two clinical situations ([Box 4.13](#)). First, it may occur in patients with leukaemia and very high white blood cell (WBC) counts; if the blood is taken and allowed to stand at room temperature without separation, the WBCs take up extracellular potassium. The second situation in which this phenomenon may occur is if blood is taken from a diabetic patient who received intravenous insulin a few minutes prior to venesection; erythrocytes take up extracellular potassium, and if blood is subsequently left to stand at room temperature unseparated for 2 h or so prior to analysis, serum potassium concentration will decrease.

Redistribution hypokalaemia in vivo. The major causes of transcellular shift of potassium occurring in vivo are also shown in [Box 4.13](#). Both alkalosis and increased plasma bicarbonate concentration without alkalosis will induce increased cellular uptake of potassium, as does endogenous or exogenous insulin. Catecholamines and β -adrenergic agonists such as adrenaline (epinephrine), salbutamol and fenoterol can all promote cellular uptake of potassium, and many patients admitted as acute medical emergencies may have significant, but transient, hypokalaemia, probably secondary to endogenous catecholamine release. Toxic chemicals such as toluene (involved in some forms of glue sniffing) and the ingestion of soluble barium salts (such as barium carbonate, which is used in certain pottery glazes, but has occasionally been mistaken for flour in cooking) have been implicated in redistribution hypokalaemia. It should be noted that barium salts used as contrast material for radiological investigations are not water soluble and do not pose a risk of toxicity.

Hypokalaemic periodic paralysis. This type of periodic paralysis occurs in two forms.

Familial hypokalaemic periodic paralysis (FHPP) is a rare autosomal dominant condition found most commonly in male Caucasians (male to female ratio 3:1 because of reduced penetrance and expression in females). The condition is characterized by attacks of flaccid paralysis affecting the limbs and trunk but rarely the facial and respiratory muscles; attacks commonly commence at night and patients present with weakness or paralysis on awakening. Attacks can last for up to 24 h and other acute manifestations of hypokalaemia such as cardiac arrhythmias may also be present. Spontaneous remission occurs with the re-establishment of normokalaemia. The attacks can begin in childhood, but often the onset may be delayed until the second decade of life. Periodicity of attacks is extremely variable between patients, with daily to yearly intervals (median four to six weeks). The attacks may be provoked by exercise followed by rest, a high carbohydrate intake, a glucose and insulin infusion, high sodium intake, adrenaline, glucagon administration and hypothermia. The total body potassium is unchanged during the attack, but potassium moves rapidly from the extracellular space into muscle. Mutations in three separate genes have now been implicated. The commonest mutations found are in the gene encoding the skeletal muscle voltage-gated calcium channel α -subunit (CACNA1S), which accounts for the majority of cases. Also described are mutations in the gene coding for the skeletal muscle voltage-gated sodium channel α -subunit (SCN4A) and the skeletal muscle voltage-gated potassium channel (KCNE3).

The management of FHPP is similar to that of any other hypokalaemia (see later), with either oral potassium supplements (up to 120 mmol/day) or, if necessary, intravenous administration, though particular care must be taken to avoid infusion of solutions containing glucose. Monitoring of serum potassium is important following the attack to ensure rebound hyperkalaemia does not occur. Prophylaxis against attacks includes a low carbohydrate diet, together with oral potassium supplements and the daily administration of spironolactone (100–200 mg), but probably most effective is the daily administration of acetazolamide (250–750 mg), although in patients with SCN4A mutations, acetazolamide may be contraindicated.

Hypokalaemic periodic paralysis with thyrotoxicosis (THPP) is a condition which primarily occurs in individuals of Chinese or Japanese extraction, but has been described in other races, including Caucasians and black people. The male preponderance is even more marked than in familial hypokalaemic periodic paralysis (FHPP) (20:1), and the age of onset is later (generally in the third decade). Links with mutations in the KCNE3 gene are described, but not in Chinese populations for whom links with mutations in CACNA1S have been made.

The clinical presentation of THPP is indistinguishable from the familial form, but the condition completely remits when the patient becomes euthyroid.

Extrarenal causes of potassium depletion. Box 4.14 lists the causes of extrarenal potassium depletion.

BOX 4.14 Causes of hypokalaemia due to extrarenal potassium depletion

Inadequate intake

- Fasting
 - Anorexia
 - During rapid cell synthesis

Increased loss

- Excessive sweating
- Gastrointestinal
 - Fistula
 - Diarrhoea
 - Cation exchange
 - Geophagia

A low dietary intake of potassium as the sole cause of hypokalaemia is rare. Renal conservation can reduce urine potassium in normal subjects to <5 mmol/L, so that without severe polyuria, a considerable lead-in period is necessary before clinically apparent hypokalaemia develops. Anorexia nervosa may, because of associated vomiting, speed the onset of clinically significant deficiency (see p. 55). One interesting cause of hypokalaemia is seen occasionally when patients with severe anaemia are treated with haematinics, resulting in a considerable and rapid increase in reticulocyte count; it has been speculated that, in certain patients, the degree of hypokalaemia that develops may induce cardiovascular death.

Excessive sweating is a potential cause of hypokalaemia. Sweat has relatively low potassium concentration (normally <10 mmol/L), but subjects who undergo severe physical exertion in hot climates, or who exercise vigorously in saunas, can lose considerable volumes of sweat (up to 12 L in a day). Subjects undergoing such ordeals may be aware of the potential for severe sodium depletion without necessarily compensating for the associated potassium depletion.

The commonest causes of extrarenal potassium depletion are those involving loss of gastrointestinal fluid rich in potassium, in particular diarrhoea. Stool water is rich in potassium with concentrations of up to 90 mmol/L, although the normal daily volume of water lost by this route results in stool potassium losses of <10 mmol/24 h. Diarrhoea of whatever cause results in an increase in stool weight and fluid content, so that severe diarrhoea may lead to stool volumes up to 2 L/24 h. Although potassium concentration in stool water tends to decrease as the volume increases, there is a limited colonic capacity for sodium/potassium exchange, and the daily losses of potassium in severe diarrhoea may exceed 100 mmol. As the stool water contains significant concentrations of bicarbonate, diarrhoea is often accompanied by a metabolic acidosis with hyperchloraemia. Thus, common causes of diarrhoea, such as bacterial causes, inflammatory bowel disease and diarrhoea associated with malabsorption syndromes, as well as less common causes, such as the watery diarrhoea-hypokalaemia achlorhydria (WDHA) syndrome – due to a pancreatic adenoma secreting vasoactive intestinal peptide (VIPoma) – are typically associated with a metabolic acidosis. Certain causes of diarrhoea, however, such as

chloride-losing diarrhoea, Zollinger–Ellison syndrome (gastrinoma) and laxative abuse, are associated with a metabolic alkalosis. Typically, these conditions are associated with hypokalaemia because of associated renal, rather than gastrointestinal, potassium loss. Villous adenoma of the rectum results in the loss of large volumes of mucus, often rich in sodium, potassium and chloride. The acid–base association can, therefore, be variable depending on predominant losses and replacement.

Potassium may be lost from the gastrointestinal tract as a result of the use of cation exchange resins such as sodium polystyrene sulphonate (Resonium A). This material is used primarily to treat clinically important hyperkalaemia. Geophagia (soil-eating) can, under certain circumstances, result in the development of hypokalaemia, as clay in the soil may bind potassium in the gut, which then passes into the stool and is lost from the body.

Renal causes of potassium depletion. The classification of renal potassium depletion adopted here is based on the associated acid–base disorder.

Renal hypokalaemic acidosis. The causes of renal hypokalaemic acidosis with normotension are summarized in Box 4.15. Acidosis, with renal potassium loss, hypokalaemia and hyperchloraemia, is found in both distal (type 1) and proximal (type 2) renal tubular acidosis (RTA), as well as in RTA induced by the carbonic anhydrase inhibitor acetazolamide.

The need for urinary diversion from the bladder, as may occur following severely impaired capacity due to fibrosis or following cystectomy for carcinoma, presents the urological surgeon with four options.

1. Bringing the ureters to the skin surface (cutaneous ureterostomy) is rarely performed, except temporarily in children: the operation will usually result in two stomata – if both kidneys are functioning – which are incontinent and require an external reservoir for collection. In addition, obstruction due to stricture is a frequent complication.
2. Transplantation of the ureters into the intact sigmoid colon (ureterosigmoidostomy) was the original diversion operation involving the gastrointestinal tract, and is occasionally still performed when medical or religious objections preclude the fashioning of an alternative exit: the urine is passed with faeces via the rectum and anus. Potentially, ureterosigmoidostomy is associated with significant metabolic complications (see below).
3. The commonest diversions now performed are those utilizing an isolated segment of bowel to form a conduit leading to the skin surface, which require a stoma bag or a constructed bladder fashioned to contain a valve to maintain continence.
4. Finally, if enough of the bladder is salvageable, then the wall may be augmented with opened segments of bowel (augmentation cystoplasty).

To understand the potential metabolic complications of potassium and other electrolytes that might arise from such procedures, it is necessary to know in some detail

BOX 4.15

Causes of renal hypokalaemia, classified by associated acid–base disorder

Renal hypokalaemic acidosis

- Renal tubular acidosis
 - Distal (type 1)
 - Proximal (type 2)
 - Carbonic anhydrase inhibitor
- Urinary tract diversion
 - Ureterosigmoidostomy
 - Colonic conduit
 - Colonic continent diversion
- Diabetic ketoacidosis

Renal hypokalaemic alkalosis with normotension

- Gastric fluid loss (chloride depleting)
- Diuretics (chloride depleting)
- Congenital chloride-losing diarrhoea
- Cystic fibrosis
- Bartter syndrome
- Gitelman syndrome

Renal hypokalaemic alkalosis with hypertension

- Primary hyperaldosteronism
- Secondary hyperaldosteronism
- Familial hyperaldosteronism types FH-1, FH-2, FH-3
- Glucocorticoid excess
- Exogenous mineralocorticoid activity
 - Congenital adrenal hyperplasia
 - 11 β -hydroxylase deficiency
 - 17 α -hydroxylase deficiency
 - Pseudohyperaldosteronism
 - Liddle syndrome
 - Apparent mineralocorticoid excess

Renal hypokalaemia without specific acid–base disorder

- Drug-related
 - Penicillin
 - Cisplatin
 - Aminoglycosides
- Leukaemia
 - Myelomonocytic leukaemia with lysozymuria
- Post relief of urinary tract obstruction
- Diuretic phase of acute kidney injury
- Magnesium depletion

how the bowel normally handles the absorption of fluid and electrolytes. The jejunum normally rapidly absorbs fluid and electrolytes, but the electrolyte components, including potassium, will be absorbed only if there is a favourable concentration gradient; if the gradient is reversed then secretion of electrolytes into the lumen can occur. In the normal ileum, active sodium absorption occurs by sodium and chloride co-transport, which is coupled to bicarbonate excretion; potassium transport is passive. In the normal colon, there is a mixture of electrogenic sodium absorption with passive chloride absorption, together with sodium and chloride co-transport coupled to bicarbonate excretion, as occurs in

the ileum. However, potassium is actively secreted into the lumen until the concentration rises to approximately 20–30 mmol/L.

As may be predicted, if bowel wall is exposed to urine, the prevalence of metabolic complications due to electrolyte redistribution will be based on the exposed surface area of bowel, the duration of contact and the type of bowel wall exposed. The diet and fluid intake will alter the urinary constituents and their concentrations, which can influence passive transport and concentration-limited active secretion. Any associated renal disease may have a direct metabolic effect. Thus, predictions in individual patients may be difficult but, in general, ureterosigmoidostomies give the greatest complication rate overall with lower rates for the other techniques. Diversions involving the sigmoid colon result in loss of bicarbonate and may result in loss of potassium, depending on the urine concentration of potassium entering the colon. Thus there is the potential for the development of significant hyperchloraemic acidosis with hypokalaemia.

Diversions involving ileal segments usually have low complication rates and, although hyperchloraemic acidosis may occur, hypokalaemia is rare. Diversions using jejunal segments are mentioned here for completeness, but the potential effect on potassium is quite different. Complications include sodium, chloride and bicarbonate depletion, with hyperkalaemia. For that reason, jejunal segments are rarely used for urinary diversion.

Diabetic ketoacidosis is associated with severe potassium depletion, but patients are usually hyperkalaemic or normokalaemic at presentation; if hypokalaemia is present, this is associated with extreme potassium depletion.

Renal hypokalaemic alkalosis. Potassium depletion with hypokalaemia and metabolic alkalosis can be further classified into normotensive and hypertensive conditions.

The causes of renal hypokalaemic alkalosis with normotension are summarized in [Box 4.15](#). Loss of gastric fluid through prolonged vomiting, or drainage of fluid via nasogastric aspiration, results in hypochloraemic alkalosis with renal excretion of bicarbonate and potassium. The associated chloride depletion results in enhanced potassium secretion by the distal tubule (see *Non-respiratory alkalosis*, Chapter 5, p. 82).

Diuretics that act primarily on the thick ascending limb of the loops of Henle (furosemide, ethacrynic acid) or on the distal collecting ducts (thiazides, chlorthalidone) can result in chloride depletion, which can further enhance the direct diuretic induced urinary potassium loss. Laxative abuse is often associated with metabolic alkalosis and renal chloride wasting; it may also be associated with simultaneous, but unrecognized, diuretic abuse or with self-induced vomiting.

Congenital chloride-losing diarrhoea is a rare condition in which ileal chloride absorption is defective. Patients present within the first decade of life with watery diarrhoea and metabolic alkalosis. The diagnosis is made from the association of hypochloraemia (serum chloride usually <90 mmol/L), alkalosis, low urine chloride concentration (<20 mmol/L), but a high stool water chloride

concentration (130–147 mmol/L). A similar clinical picture is seen in some patients with Zollinger–Ellison syndrome and in the rare condition of systemic mastocytosis secondary to basophil leukaemia; in the latter condition there is hypersecretion of gastric fluid due to histamine release, with diarrhoea, intestinal loss of chloride and secondary renal loss of potassium.

Rarely, cystic fibrosis (CF) may result in hypokalaemia in association with hypochloraemic alkalosis because of the high sweat chloride concentration in CF.

Barter syndrome, a recessive autosomal condition, presents with hypokalaemic alkalosis in association with hyperreninaemic hyperaldosteronism. There is renal wasting of potassium and chloride, but patients are resistant to the pressor effects of angiotensin II and, thus, are normotensive or mildly hypotensive. Patients present in childhood with failure to thrive, unexplained hypokalaemia and, occasionally, renal impairment. Renal biopsy typically reveals hyperplasia of the juxtaglomerular apparatus. Occasionally, presentation is delayed until adulthood, at which time the major differential diagnosis is of diuretic abuse.

Barter syndrome is caused by a loss of function mutation in one of the genes coding for three transport proteins located in the epithelial cells of the thick ascending limb of the loops of Henle – respectively the $\text{Na}^+\text{-K}^+\text{-2Cl}^-$ co-transporter (NKCC2), the potassium-secreting channel (ROMK1), which are both located on the apical membrane and the basolateral chloride channel (CLC-K6). Variations in phenotypic expression of Barter syndrome have been explained by association with specific transport protein mutations, for example early presentation, associated hypercalciuria and impaired urine-concentrating ability are typically associated with apical transport proteins, whereas later presentation and normal urine calcium excretion are associated with CLC-K6 mutations.

Gitelman syndrome is an autosomal recessive condition presenting with hypokalaemic alkalosis, but, unlike Barter syndrome, is associated with hypomagnesaemia and hypocalciuria. Gitelman syndrome is caused by loss of function mutations in the gene coding for the thiazide-sensitive $\text{Na}^+\text{-Cl}^-$ co-transporter (NCCT), situated in the distal convoluted tubules. This disorder may be asymptomatic in childhood and then present in adult life with weakness, fatigue, paraesthesia and, rarely, tetany.

A detailed classification of conditions causing renal hypokalaemic alkalosis with hypertension is given earlier in this chapter (see [Box 4.4](#), p. 37), but these are summarized in [Box 4.15](#).

The association of hypokalaemic alkalosis with hypertension is due to functional mineralocorticoid excess, resulting in sodium retention and distal tubular potassium wasting with net acid excretion. Excess production of aldosterone may be primary (hyporeninaemic) due to adenoma, hyperplasia or, rarely, carcinoma of the zona glomerulosa of the adrenal cortex (see Chapter 18). Three forms of familial hyperaldosteronism are currently described. Familial hyperaldosteronism type I (FH-I), a rare autosomal dominant condition, is also known as glucocorticoid suppressible hyperaldosteronism. It is due

to genetic recombination, resulting in the adrenocorticotrophic hormone (ACTH)-responsive region of 11-hydroxylase being incorporated into aldosterone synthase (hybrid CYP11B1/CYP11B2 gene mutation). FH-II is not glucocorticoid suppressible, appears to be up to five times more prevalent than FH-I and is clinically indistinguishable from apparent sporadic primary hyperaldosteronism. Family studies reveal linkage at chromosome 7p22. Candidate genes involved in cell cycle control are implicated as adrenal hyperplasia and adrenal adenomas are reported as common in FH-II patients. FH-III, also not glucocorticoid suppressible, has been described in a single affected family in which all three members had hypertension, hyperaldosteronism and high plasma concentrations of the hybrid steroids 18-hydroxycortisol and 18-oxocortisol. Only bilateral adrenalectomy could normalize blood pressure.

Secondary (hyper-reninaemic) hyperaldosteronism with hypertension is seen in patients with accelerated (malignant) hypertension and renal artery stenosis. Renin-producing tumours, such as benign haemangiopericytoma of the juxtaglomerular apparatus, are rare causes of hyper-reninaemic hyperaldosteronism.

Cortisol possesses only weak mineralocorticoid activity, but when plasma concentrations are sufficiently elevated, a mineralocorticoid effect results. Cushing syndrome, ectopic ACTH syndrome or the administration of exogenous glucocorticoid with mineralocorticoid activity can result in significant hypokalaemic alkalosis and hypertension.

Two forms of congenital adrenal hyperplasia result in excess production of deoxycorticosterone (DOC), which has mineralocorticoid action. These are 11 β -hydroxylase deficiency and 17 α -hydroxylase deficiency.

Pseudohyperaldosteronism can present in two forms. Liddle syndrome is a rare autosomal dominant condition that clinically resembles primary hyperaldosteronism, but in which aldosterone and renin are suppressed. The onset of hypertension and hypokalaemia occurs typically in childhood, but, even within a single family, penetrance can be extremely variable. The condition is caused by mutation within the gene coding for the collecting duct amiloride-sensitive epithelial sodium channel (ENaC), resulting in an 'open' channel, unregulated by the mineralocorticoid receptor (compared with loss of function mutations causing pseudohypoaldosteronism type 1, see p. 35).

Apparent mineralocorticoid excess has been described in children and, as in Liddle syndrome, there is hyporeninaemic hypoaldosteronism, but unlike Liddle syndrome, patients are responsive to spironolactone. This rare autosomal recessive condition is caused by impaired activity of the enzyme 11 β -hydroxysteroid dehydrogenase (11 β -HSD) due to mutations within the HSD11B2 gene. This enzyme normally inactivates cortisol in the kidney by conversion to cortisone (see p. 29). The mineralocorticoid receptor has equal affinity for cortisol, and selectivity for aldosterone is normally maintained by 11 β -HSD activity. An acquired form of this syndrome can be produced by excess ingestion of glycyrrhizic acid found in natural liquorice root, as well as glycyrrhetic acid, the active component of the peptic ulcer drug carbenoxolone.

Renal hypokalaemia without specific acid-base disorder. There remains a miscellaneous group of conditions that result in renal loss of potassium, but without a specific acid-base disorder: these conditions are summarized in [Box 4.15](#).

Several drugs can result in renal potassium loss. Penicillins, especially those administered in large molar quantities, e.g. carbenicillin, are excreted in anionic form in the urine and are associated with kaliuresis. Both cisplatin and the aminoglycoside group of antibiotics have been associated with hypokalaemia, probably because of associated magnesium depletion (see below).

Acute myelomonocytic leukaemia is associated with renal loss of potassium, especially in association with renal excretion of lysozyme released from the leukaemic cells.

Both the diuretic phase of acute kidney injury and that following the relief of urinary tract obstruction can result in a negative potassium balance with associated hypokalaemia.

Finally, there is increasing evidence that magnesium depletion, of whatever aetiology, is associated with an increase in renal potassium loss. Intracellular magnesium deficiency affects potassium excretion by reducing Na⁺,K⁺-ATPase activity proximally within the nephron, thus increasing distal sodium delivery and hence potassium secretion. In addition, magnesium deficiency reduces magnesium dependent inhibition of ROMK channels within the collecting duct principal cells, which also increases potassium secretion. In patients who are both potassium and magnesium depleted, supplementation with potassium will not result in correction of hypokalaemia unless magnesium is also supplemented.

Laboratory investigation of hypokalaemia

In the majority of patients presenting with hypokalaemia, the cause is clear from the clinical history. The role of the laboratory is then to provide a monitoring service for serum potassium and no further specific investigation is required. However, on occasion, patients presenting with inexplicable, consistent, severe hypokalaemia provide a considerable diagnostic challenge.

The first stage in differentiating the cause of hypokalaemia is to ensure that the serum potassium is a true reflection of in vivo concentration and not due to in vitro redistribution (see [Box 4.13](#)). A fresh sample of blood should be obtained for immediate separation and measurement of sodium and potassium concentration, and at the same time, renal function should be assessed by the measurement of serum urea or creatinine concentrations (or both), and magnesium should be measured to exclude concomitant magnesium depletion. If possible, blood should also be obtained for assessment of acid-base status. The clinical history, including a detailed past and present drug history, is essential. The next step in laboratory investigation is to differentiate renal from extrarenal causes of hypokalaemia by measuring urine potassium. The definition of renal potassium conservation is not absolute, but, in the presence of hypokalaemia,

a 24h potassium excretion of <15 mmol and a random urine potassium/creatinine ratio of <1.5 mmol/mmol provide strong evidence of extrarenal loss. The trans-tubular potassium gradient (TTKG) can also be used. TTKG is calculated by dividing the urine/serum potassium concentration by the urine/serum osmolality ratio (see [Appendix 4.1h](#)). The TTKG in renal conservation of potassium is expected to be <2.

When potassium is not conserved by the kidneys, the presence of a metabolic acidosis would suggest one of the causes listed in [Box 4.15](#) under renal hypokalaemic acidosis, although the majority of such conditions, excepting certain RTAs, should be apparent from the clinical history.

Hypokalaemia with metabolic alkalosis can be further differentiated by measurement of urine chloride. Again, no absolute criteria are available, but a random urine chloride <20 mmol/L (or, with polyuria, 2 mmol/mmol creatinine) indicates renal chloride retention and implies chloride depletion. Those conditions listed in [Box 4.15](#) as chloride-depleting include certain conditions that frequently provide a diagnostic challenge, such as diuretic abuse and surreptitious vomiting. If the urine chloride concentration exceeds 20 mmol/L, then Bartter syndrome, Gitelman syndrome and diuretic abuse (with recent diuretic ingestion) should be considered; in the presence of hypertension, those conditions associated with glucocorticoid or mineralocorticoid excess should be considered, although it should be remembered that hypertension is common, but mineralocorticoid-induced hypertension is not. (For full investigation of states of glucocorticoid or mineralocorticoid excess, see Chapters 18 and 38). Particular care should be exercised when interpreting urine electrolyte concentrations in hypokalaemic states. As mentioned previously, the concentration of potassium or chloride may be reduced if polyuria secondary to hypokalaemia is present. Therefore, it may be necessary to express concentrations in terms of creatinine or some other measure of urine concentration. The excretion or retention of electrolytes may be time-dependent; for example, diuretics may induce a kaliuresis and chloururesis following ingestion, which reverts between doses to a kaliuresis with chloride retention. If stopping the diuretic, with subsequent chloride, but not potassium replacement, induces urine potassium conservation, the implication is that there is extrarenal loss of potassium. In addition, two or more conditions may coexist, for example anorexia and vomiting, laxative abuse and diuretic abuse. Finally, severe potassium depletion may of itself induce renal chloride wasting, thus potentially, a severe extrarenal loss of potassium may result in renal potassium wasting.

Management of hypokalaemia

The decision to treat hypokalaemia will depend on the severity of the condition, the presence of symptoms, particularly muscle weakness, electrocardiographic evidence of cardiovascular effects and concomitant therapies, for example with cardiac glycosides, which increase cardiac sensitivity to hypokalaemia.

TABLE 4.6 Potassium content of oral preparations

Preparation	Potassium content
Salt	(mmol/g)
Potassium chloride	13.4
Potassium bicarbonate	9.9
Potassium citrate	9.2
Fruit juice	(mmol/100 mL)
Tomato juice	8.2
Orange juice	3.0
Grapefruit juice	3.0
Apple juice (and craft cider)	3.2

The magnitude of any potassium deficit can only be estimated. For hypokalaemia associated with a deficit rather than redistribution, a rough guide is that a serum potassium of 3.0 mmol/L equates with a total body deficit of 300 mmol and a serum potassium of 2.0 mmol/L, with a total deficit of 700–800 mmol.

Unless hypokalaemia is severe and potentially life-threatening (serum potassium <2.5 mmol/L), then in general, oral replacement is preferred. Although certain fruit juices contain relatively high concentrations of potassium, the volumes required to provide sufficient potassium for replacement may be impracticable (see [Table 4.6](#)). Potassium salts are available in various formulations, including syrups, effervescent preparations and sustained-release tablets. Potassium chloride is usually the preferred salt, and potassium bicarbonate should only be used in hyperchloraemic states. Potassium citrate is usually prescribed to alkalinize the urine and reduce the discomfort of urinary tract infections, but can be substituted for potassium bicarbonate in the treatment of hypokalaemia. Currently, the preferred sustained-release formulation is a microencapsulated form, which spreads the release of the potassium salt within the gastrointestinal tract and minimizes the risk of ulceration and stricture formation. Oral replacement is usually prescribed in doses that equate with a normal adult dietary intake of between 40 and 120 mmol/day, although up to 200 mmol/day may be required and tolerated.

In states of severe hypokalaemia, or when oral replacement is not possible, potassium can be given intravenously into peripheral veins. The maximum rate of infusion should not usually exceed 20 mmol/h, although this limit can, with cardiac monitoring, be increased to 40 mmol/h. The maximum concentration of potassium in the infusion should not usually exceed 40 mmol/L. If it is necessary to minimize fluid intake, or if hypokalaemia proves resistant to such replacement, concentrations up to 80 mmol/L may be infused into a central vein.

Hyperkalaemia

There is no strict definition of hyperkalaemia, but persistent serum potassium concentrations >5.3 mmol/L (plasma >5.0 mmol/L) warrant further investigation. The most significant clinical effect of hyperkalaemia is on cardiac function, most importantly to cause cardiac

arrest. This risk does not become appreciable until serum potassium is $>6\text{mmol/L}$, but is considerable at concentrations $>8\text{mmol/L}$, particularly if the increase is rapid. Electrocardiographic changes become apparent at lower concentrations of potassium and are more prominent if the hyperkalaemia is associated with hypocalcaemia, hyponatraemia, hypermagnesaemia or acidosis. Occasionally, patients present with a form of ascending muscular weakness resembling the Guillain-Barré syndrome. [Box 4.16](#) shows the major clinical effects of hyperkalaemia.

Causes of hyperkalaemia

Hyperkalaemia may reflect an inappropriate retention of potassium within the body or an alteration in the distribution of potassium intra- and extracellularly. This alteration in distribution may occur *in vivo*, or *in vitro*, when there has been redistribution of potassium from blood cells to serum during the time between venesection and separation of serum.

Redistribution hyperkalaemia *in vitro*. This form of hyperkalaemia is sometimes known as spurious or pseudohyperkalaemia ([Box 4.17](#)). The most common form is due simply to *in vitro* loss of potassium from erythrocytes, and may or may not be associated with visible haemolysis. This type of spurious increase in serum potassium is also seen in certain patients with leukaemia with extremely high WBC counts ($>100\times 10^9/\text{L}$), and in patients with thrombocythaemia when platelet counts exceed $1000\times 10^9/\text{L}$; in both situations the cells are fragile and lyse during blood clotting. To exclude these causes in patients with leukaemia or thrombocythaemia,

BOX 4.16 The clinical effects of hyperkalaemia

Cardiovascular

- Electrocardiographic changes
 - Tall T waves
 - Prolonged PR interval
 - Flat P waves
 - Widening QRS complex
 - 'M' complex sine wave
 - Ventricular fibrillation or asystole

Neuromuscular

- Paraesthesia
- Weakness, flaccid paralysis

Renal

- Natriuresis
- Decreased renal ammoniogenesis
- Decreased reabsorption of bicarbonate in the proximal tubule

Endocrine

- Decreased renin secretion
- Increased aldosterone secretion
- Increased insulin secretion

BOX 4.17 Causes of redistribution hyperkalaemia

In vitro redistribution

- Haemolysis of erythrocytes
- Release from white blood cells (high leukaemic counts)
- Release from platelets (thrombocythaemia)
- Storage of whole blood at low temperature
- Familial pseudohyperkalaemia

In vivo redistribution

- Acidosis
- Insulin deficiency with hyperglycaemia
- Drugs
 - Succinylcholine
 - β -Adrenergic blockers
- Acute tissue damage
- Hyperkalaemic periodic paralysis

blood should be collected into a heparinized container and the plasma separated quickly from the cells at room temperature.

Whole blood samples stored at 4°C will eventually release potassium from red cells into the plasma, without necessarily any evidence of haemolysis. In certain families, this tendency is greatly accentuated, giving rise to the condition familial pseudohyperkalaemia, which is due to disordered cation transport in the erythrocyte membrane.

Redistribution hyperkalaemia *in vivo*. The major causes of transcellular shift of potassium occurring *in vivo* are also shown in [Box 4.17](#). As previously discussed, acidosis, particularly that induced by mineral acid, will result in increased extracellular potassium concentration. Acidosis and hypertonicity due to hyperglycaemia and insulin deficiency lead to the hyperkalaemia frequently seen in potassium-depleted diabetic patients with ketoacidosis.

Various drugs may promote or accentuate hyperkalaemia. Succinylcholine, a depolarizing muscle relaxant, results in some increase in plasma potassium in all subjects, but this is particularly evident in patients with increased total body potassium. Thus succinylcholine should be avoided in all patients who are hyperkalaemic. β -Adrenergic blocking agents have been associated with minor increases in plasma potassium concentration, but this may be greatly accentuated during vigorous exercise.

The release of potassium from cells during strenuous exercise is well described, and a similar phenomenon may occur locally in an ischaemic limb, such as may happen during prolonged venous stasis prior to venesection. Hyperkalaemia has been described in patients undergoing chemotherapy for malignancy, which causes massive lysis of neoplastic cells. This phenomenon has been described in the treatment of chronic lymphocytic leukaemia, acute lymphoblastic leukaemia and lymphosarcoma. Its occurrence emphasizes the need for maintenance of hydration and for careful electrolyte monitoring during aggressive chemotherapy. Acute

haemolytic disorders can give rise to hyperkalaemia by a similar mechanism.

Hyperkalaemic periodic paralysis. Hyperkalaemic periodic paralysis (HYPP) is a rare autosomal dominant condition that presents with attacks of muscular weakness, paralysis (usually sparing the respiratory muscles) and an associated acute rise in plasma potassium (up to 8.0 mmol/L). The periodicity of attacks is variable, ranging from daily to only a few attacks a year. This type of periodic paralysis is provoked by a high potassium intake, glucocorticoids, hypothermia and during the recovery phase after vigorous exercise. During attacks, electrocardiographic monitoring shows tall T waves but cardiac arrhythmias are rare. The disorder is due to gain of function mutations in the skeletal muscle voltage-gated sodium channel X subunit (SCN4A) gene. The management of HYPP is administration of the β_2 -agonist salbutamol, which can easily be taken by inhalation. Salbutamol has also been used prophylactically.

Potassium retention. The major causes of potassium retention are listed in Box 4.18. An increased potassium intake alone, without associated renal impairment, in practice only occurs as an iatrogenic complication of inappropriate intravenous loading. Excessively high oral loads of potassium are counteracted by a combination of reduced gastrointestinal absorption, vomiting and diarrhoea.

A reduction in glomerular filtration rate (GFR), of whatever cause, increases the risk of development of hyperkalaemia. However, on a normal diet containing approximately 100 mmol of potassium per day, GFR may fall to less than 10 mL/min before a significant risk of hyperkalaemia arises, and increased colonic secretion of potassium may further protect the patient when GFR is reduced further. It is important, however, that patients with significantly reduced GFR avoid high intakes of potassium or any condition that results in endogenous shifts of potassium.

Decreased tubular secretion of potassium may occur in response to treatment with the potassium-sparing diuretics: spironolactone, triamterene or amiloride. Hyperkalaemia is a particular risk when such a drug is used in a patient with impaired renal function or in a patient with a high potassium intake. Urinary diversions that involve jejunal segments (see p. 55) result in a significant incidence of hyperkalaemia due to reabsorption

BOX 4.18 Causes of hyperkalaemia due to potassium retention

Increased intake

- Increased post-tubular reabsorption following surgical urinary diversion involving jejunal segments

Decreased output

- Decreased glomerular filtration rate
- Decreased tubular secretion
 - Potassium-sparing diuretics
 - Syndromes of hypoaldosteronism

of potassium from the urine while contained within the segment.

Syndromes of hypoaldosteronism. This broad group of conditions is summarized in Table 4.7. Primary hyperreninaemic hypoaldosteronism is found in Addison disease (see Chapter 18) and the rare condition corticosterone methyl oxidase deficiency, in which aldosterone synthesis is impaired. Heparin given in continuous high doses can also inhibit aldosterone synthesis through inhibition of 18-hydroxylase and possibly by inducing atrophy in the zona glomerulosa.

Two forms of congenital adrenal hyperplasia (CAH) result in impaired mineralocorticoid synthesis. 21-Hydroxylase deficiency is the most common form of CAH. Virilization is characteristic, but not all patients exhibit features of hypoaldosteronism; up to two-thirds develop renal sodium wasting with hyperkalaemia. 3 β -Hydroxydehydrogenase deficiency is a rarer condition, in which the majority of patients have renal sodium wasting and hyperkalaemia.

Hypoaldosteronism may occur in patients receiving angiotensin-converting enzyme (ACE) inhibitors such as captopril. Patients at particular risk of developing hyperkalaemia include those with associated renal impairment or with high renin concentrations, for example resulting from congestive heart failure or renal artery stenosis.

Hyporeninaemic hypoaldosteronism is an increasingly recognized syndrome in which patients present with hyperkalaemia out of proportion to a reduction in GFR. The majority of such patients develop a hyperchloraemic metabolic acidosis and fall, therefore, into the classification of type 4 renal tubular acidosis (see Chapter 5). The syndrome is particularly prevalent in elderly patients with type 2 diabetes, but is also found in many other diseases in association with interstitial nephritis, including systemic lupus erythematosus, multiple myeloma, chronic obstructive uropathy, gout, sickle cell disease, lead nephropathy, following renal transplantation and in association with treatment with prostaglandin

TABLE 4.7 Syndromes of hypoaldosteronism with relative plasma renin activities and aldosterone concentrations (\uparrow , increased; \downarrow , decreased; N, normal)

Syndrome	Renin	Aldosterone
Primary hypoaldosteronism		
Addison disease	\uparrow	\downarrow
Isolated aldosterone deficiency	\uparrow	\downarrow
Heparin treatment	\uparrow	\downarrow
Congenital adrenal hyperplasia	N to \uparrow	N to \downarrow
Angiotensin-converting-enzyme inhibition	N to \uparrow	N to \downarrow
Hyporeninaemic hypoaldosteronism	\downarrow	\downarrow
Secondary tubular disorders	N	N
Pseudohypoaldosteronism		
type I	\uparrow	\uparrow
type II	\downarrow	N to \downarrow

synthetase inhibitors and treatment with ciclosporin. Plasma renin activity is reduced, as is aldosterone concentration. Renin response to upright posture or to salt depletion is also reduced. The pathogenesis is not fully understood, but includes structural damage to the kidney, including the juxtaglomerular apparatus, and there is evidence in diabetes of impaired conversion of renin precursor to active renin. One intriguing aspect of hyporeninaemic hypoaldosteronism is why hyperkalaemia occurs, as aldosterone secretion is known to be stimulated by hyperkalaemia. However, it is likely that this direct stimulation of aldosterone can only occur in the presence of angiotensin II. An important aspect of this syndrome is that, because it occurs more frequently in the elderly, drugs that suppress both renin and aldosterone, such as β -blockers and calcium channel blockers as well as prostaglandin synthetase inhibitors, should generally be avoided in this age group.

Interstitial nephritis may give rise to hyperkalaemia in which aldosterone secretion is not suppressed. The pathogenesis in these conditions is presumed to be a direct impairment of tubular secretion of potassium. Many of the diseases associated with tubular dysfunction are those that are also associated with the development of hyporeninaemic hypoaldosteronism and include obstructive uropathy, sickle cell disease and systemic lupus erythematosus. The main clinical difference between direct tubular dysfunction and hyporeninaemic hypoaldosteronism is the failure of the former condition to respond adequately to mineralocorticoid replacement, and alternative treatments such as thiazide diuretics must be used.

Pseudohypoaldosteronism (PHA) is a term applied to a heterogeneous group of rare disorders linked by association with hyperkalaemia, metabolic acidosis, normal glomerular function and apparent renal tubular unresponsiveness or resistance to mineralocorticoids. The conditions are broadly segregated into PHA types I and II.

Pseudohypoaldosteronism type I can be inherited either as an autosomal dominant or autosomal recessive trait. Pseudohypoaldosteronism type II, alternatively known as Gordon syndrome, is inherited as an autosomal dominant condition. The major characteristics of these syndromes are summarized in Table 4.8.

Laboratory investigation of hyperkalaemia

The first stage in any laboratory investigation of hyperkalaemia is to ensure that the serum potassium is a true reflection of in vivo concentration. The storage of whole blood specimens in a refrigerator at 4°C is widely practised by clinicians in the belief that this will aid the preservation of a specimen, but this practice can greatly increase serum potassium without any evidence of haemolysis. This effect is particularly noticeable in familial pseudohyperkalaemia, but can affect any blood sample if storage is prolonged for 8–12 h. As previously mentioned, hyperkalaemia in patients with high WBC counts or platelet counts should be confirmed in freshly separated plasma, rather than serum.

Having confirmed true hyperkalaemia, the clinical history is required with particular emphasis on drug and dietary regimens, with information being sought regarding possible causes of in vivo redistribution.

Assessment of blood acid–base status and glucose concentration may be valuable. Serum potassium is usually measured with other electrolytes and markers of renal function, including urea and creatinine. When GFR is reduced to <10 mL/min, hyperkalaemia is likely to develop unless dietary potassium is restricted. The measurement of urine potassium output is of marginal value except possibly in steady-state conditions when a 24 h urine potassium will provide evidence of excessive ingestion.

If GFR is not reduced sufficiently to explain the hyperkalaemia, a syndrome of hypoaldosteronism should be considered. The measurement of TTKG may provide

TABLE 4.8 Characteristics of pseudohypoaldosteronism (PHA) types I and II

	PHA type I		PHA type II
	Autosomal dominant PHA	Autosomal recessive PHA	Autosomal dominant Gordon syndrome
Age at presentation	Newborn/early infancy	Newborn/early infancy	Late childhood/adulthood
Mechanism	Loss of function mineralocorticoid receptor mutation <i>NR3C2</i> gene	Loss of function epithelial sodium channel (ENaC) <i>SCNN1A</i> , <i>SCNN1B</i> , <i>SCNN1G</i> subunit genes	Enhanced distal tubular chloride reabsorption in preference to potassium excretion <i>WNK1</i> , <i>WNK4</i> genes
Organ affected	Kidneys	Kidneys, sweat glands, salivary glands, colon	Kidneys
Blood pressure	Hypotension	Hypotension	Usually hypertension
Renal sodium loss	Present	Present	Absent
Treatment	Sodium supplements Restricted potassium intake K ⁺ -binding resins	Sodium supplements Restricted potassium intake K ⁺ -binding resins High-dose fludrocortisone or carbenoxolone	Dietary sodium restriction Thiazide diuretics
Prognosis	Improvement with advancing age	Lifelong therapy required	Lifelong therapy required

useful information (see [Appendix 4.1h](#)). The expected physiological response in hyperkalaemia would result in a TTKG of >10. A value for TTKG of <7 is compatible with hypoaldosteronism. Normally, the clinical presentation of Addison disease is sufficiently characteristic to require only studies of cortisol response to exogenous ACTH preparations, but other forms of hypoaldosteronism require measurement of renin and aldosterone. Unfortunately, the various combinations of findings do not fit snugly with clinical classification of disorders. Hyper-reninaemic hypoaldosteronism is found in primary hypoaldosteronism, the exceedingly rare condition of corticosterone methyl oxidase deficiency, some forms of CAH, and with ACE inhibitor treatment. Hyporeninaemic hypoaldosteronism is found in the conditions grouped under the syndrome of the same name and is also found in the rare PHA type II (Gordon syndrome). In contrast, high renin and high aldosterone are found in PHA type I. However, normal renin concentrations and aldosterone activity may be found in those conditions causing interstitial nephritis with direct tubular inhibition of potassium excretion. Renin and aldosterone values are not always easy to interpret, as high plasma potassium directly stimulates aldosterone and suppresses renin activity. In addition, aldosterone concentration increases as the GFR falls without a corresponding change in renin activity. In situations in which diagnostic difficulty may exist, for example in hyporeninaemic hypoaldosteronism, it may be necessary to reassess aldosterone concentrations when plasma potassium has been reduced to within the reference range; alternatively a trial of mineralocorticoid replacement will differentiate hyporeninaemic hypoaldosteronism from direct tubular dysfunction.

Management of hyperkalaemia

Hyperkalaemia, particularly when severe (>6.0 mmol/L), is a serious condition requiring immediate treatment because of the risk of sudden death. Emergency management is intravenous 10% calcium gluconate – 10 mL injected over 60–120 s and repeated every 15 min or so until the electrocardiographic changes improve (maximum dose 50 mL). This does not correct the hyperkalaemia, but is directly cardioprotective. In patients receiving digoxin, calcium gluconate should be infused more slowly – 10 mL over 30 min – to avoid digoxin toxicity induced by hypercalcaemia.

Two therapeutic regimens are available to lower plasma potassium rapidly. Glucose (50 mL of 50% glucose) can be infused over 15 min together with ten units of soluble insulin. This regimen may be repeated at hourly intervals and should be accompanied by serum potassium and plasma glucose monitoring. Alternatively, plasma potassium may be reduced by the infusion of 50–100 mL of 4.2% sodium bicarbonate (500 mmol/L) over a 15–30 min period (provided that the patient is not sodium overloaded).

For haemodialysis patients with hyperkalaemia, a regimen that has been shown to be of equal benefit to glucose and insulin in reducing plasma potassium is

treatment with the β_2 -adrenergic agonist salbutamol. Inhalation of nebulized salbutamol (10–20 mg) will reduce plasma potassium by approximately 1 mmol/L within 30 min.

If hyperkalaemia is the result of increased body stores of potassium, then this excess must be removed. The relationship of plasma potassium to excess body potassium is highly variable and is not reliably predictable. As an approximation, elevation of plasma potassium by 1 mmol/L above normal without evidence of redistribution will roughly equate with a 200 mmol total excess. Polystyrene sulphonate resins (sodium or calcium salts) may be given by mouth (15 g, 3–4 times daily in water) or as an enema (30 g in methylcellulose, retained for 9 h). As an approximation, each gram of resin removes 1 mmol of potassium, so that up to 60 mmol may be removed in each 24 h period. For more dramatic reductions in body potassium, renal replacement treatment is required. Peritoneal dialysis is capable of removing 10–15 mmol of potassium each hour, whereas the efficiency of haemodialysis will allow up to 30 mmol to be removed each hour.

If a reversible cause of hyperkalaemia cannot be identified, then therapy for chronic hyperkalaemia should be designed to minimize the recurrence of severe hyperkalaemia. This should include a reduction in dietary potassium intake (<50 mmol/day) and avoidance of volume contraction and drugs that cause redistribution hyperkalaemia. Sodium bicarbonate and thiazide diuretics may be useful and, in some cases, treatment with mineralocorticoids.

CONCLUSION

The physiological control over sodium, water and potassium within the human body is a complex interrelated series of systems of extreme precision and sensitivity. These systems regulate the extracellular fluid volume, the extra- and intracellular solute content, the intracellular volume and neuromuscular function, and, therefore, indirectly influence myriad functional and metabolic processes essential for life. The pathological causes and consequences of recognized abnormalities in the control of sodium, water and potassium have been explored in this chapter, together with the details of diagnosis and treatment.

Further reading

Adrogue HJ, Madias NE. Hyponatremia. *N Engl J Med* 2000;342:1581–9.

Focusing on the quantitative approach to rational treatment of hyponatraemia.

Bartter FC, Schwartz WB. The syndrome of inappropriate secretion of antidiuretic hormone. *Am J Med* 1967;42:790–806.

A medical classic by the authors who originally coined the term 'syndrome of inappropriate antidiuretic hormone secretion' (SIADH). The diagnostic criteria for this syndrome are clearly defined.

Ball SG. Vasopressin and disorders of water balance: the physiology and pathophysiology of vasopressin. *Ann Clin Biochem* 2007;44:417–31.

A review of the disorders of water balance.

Greenberg A, Leich RW. Treatment of chronic hyponatremia: now we know how, but do we know when or if? *J Am Soc Nephrol* 2010;21:552–5.

Editorial comment on the appropriate usage of vasopressin receptor antagonists (vaptans) in chronic hyponatraemia

Halperin ML, Kamel KS. Potassium. *Lancet* 1998;352:135–40.

A general review of hyper- and hypokalaemia. Rational approaches to diagnosing the underlying causes are explored, and management algorithms presented.

King L, Yasui M, Agre P. Aquaporins in health and disease. *Mol Med Today* 2000;6:60–5.

The history of the discovery of these water-channel proteins and their physiology and pathology.

Kumar S, Berl T. Sodium. *Lancet* 1998;352:220–8.

A general review of hypo- and hypernatraemia with particular emphasis on the treatment of acute and chronic hyponatraemia.

Penney MD, Walters G. Are osmolality measurements clinically useful? *Ann Clin Biochem* 1987;24:566–71.

A critical review of the clinical usefulness of body fluid osmolality measurements, including measurements applied to fluid balance disorders.

Robertson GL, Aycinena P, Zerbe RL. Neurogenic disorders of osmoregulation. *Am J Med* 1982;72:339–53.

The osmoregulatory control of AVP in hypodipsic hypernatraemic conditions is discussed in detail.

Sayer JA, Pearce SHS. Diagnosis and clinical biochemistry of inherited tubulopathies. *Ann Clin Biochem* 2001;38:459–70.

A review covering in detail disorders caused by inherited tubulopathies. Disorders affecting renal sodium and potassium handling are included.

Unwin RJ, Luft FC, Shirley DG. Pathophysiology and management of hypokalaemia: a clinical perspective. *Nat Rev Nephrol* 2011;7:75–84.

A review covering in detail the pathophysiology of hypokalaemia and its treatment.

Waters P, Hack MA, Richards J, Penney MD. Quantitating nocturia: a study into the recording of solute and water excretion to determine causation. *Ann Clin Biochem* 2011;48:321–6.

Outlining an approach to the investigation of nocturnal polyuria.

Zerbe RL, Stropes L, Robertson GL. Vasopressin function in the syndrome of inappropriate antidiuresis. *Annu Rev Med* 1980;31:315–27.

Patterns of plasma arginine vasopressin (AVP) responses to hypertonic saline infusion in patients with dilutional hyponatraemia are presented.

APPENDIX 4.1 FORMULAE

Formulae (a)–(e) provide approximations only and are supplied as illustrative guidelines. Any corrective procedure based upon any of these formulae should be accompanied by detailed clinical and laboratory monitoring.

[Concentrations]: serum (s), plasma (p) or urine (u) in mmol/L or mmol/kg (osmolality).

(a) Estimate of reduction in ECF volume from rise in haematocrit (HCT) when no blood loss has occurred

$$\text{ECF volume reduction (litres)} = 0.2 \times \text{body weight (kg)} \times \left[1 - \left[\frac{\text{normal hct}}{\text{measured hct}} \right] \right]$$

(b) Estimate of sodium deficit in patients with hypovolaemic hyponatraemia

$$\text{sodium deficit (mmol)} = 0.6 \times \text{body weight (kg)} \times (140 - [\text{Na}^+]_s)$$

(c) Estimate of water deficit in hypernatraemia

$$\text{water deficit (litres)} = 0.6 \times \text{body weight (kg)} \times \left[1 - \frac{140}{[\text{Na}^+]_s} \right]$$

(d) Estimate of the expected sodium depression in hyperglycaemia-induced hyponatraemia

$$\text{expected sodium depression (mmol/L)} = 0.288 \left([\text{glucose}]_p - 5.5 \right)$$

Reference

Adapted from Katz MA. Hyperglycemia-induced hyponatremia – calculation of expected serum sodium depression. *N Engl J Med* 1973;289:843–4.

(e) Calculation of serum osmolality

$$\begin{aligned} \text{calculated osmolality} &= 1.89 [\text{Na}^+]_s + 1.38 [\text{K}^+]_s \\ &+ 1.03 [\text{urea}]_s + 1.08 [\text{glucose}]_p \\ &+ 7.45 \end{aligned}$$

Reference

Adapted from Behagat CI, Garcia-Webb P, Fletcher E, Beilby JP. Calculated vs measured osmolalities revisited. *Clin Chem* 1984;30:1703–5.

(f) Calculation of osmolal gap

$$\text{osmolal gap} = \text{measured osmolality} - \text{calculated osmolality}$$

(g) Estimate of sodium required in acute water intoxication

(To increase serum sodium to 125 mmol/L):

$$\text{sodium required (mmol)} = (125 - [\text{Na}^+]_s) \times 0.6 \times \text{body weight (kg)}$$

(Hypertonic saline (5%) = 855 mmol/L)

(h) Calculation of the transtubular potassium gradient (TTKG)

$$\text{TTKG} = \frac{[\text{K}^+]_u \times [\text{Osm}]_s}{[\text{K}^+]_s \times [\text{Osm}]_u}$$

Reference

Adapted from Ethier JH, Kamel KS, Magner PO, Lemann Jr J, Halperin ML. The transtubular potassium concentration in patients with hypokalemia and hyperkalemia. *Am J Kidney Dis* 1990;15:309–15.

APPENDIX 4.2 DYNAMIC FUNCTION TESTS

(a) Water deprivation test

This test is used to differentiate between cranial diabetes insipidus, nephrogenic diabetes insipidus or primary polydipsia as causes of polyuria.

The patient is denied fluid and sloppy food from 08.30 h onwards. All patients should be observed closely to prevent covert access to fluid. The following protocol should be followed. Urine should, if possible, be collected hourly, the volume recorded and an aliquot tested for osmolality. Blood should be obtained for serum sodium and osmolality measurements. Accurate weight recordings are required to monitor loss and, in certain cases, detect surreptitious fluid consumption.

Time of urine collection (h)	Time of blood collection (h)	Time of weighing (h)
09.00–10.00	09.00	09.00
10.00–11.00	–	–
11.00–12.00	–	–
12.00–13.00	12.00	12.00
13.00–14.00	–	–
14.00–15.00	–	14.00
15.00–16.00	–	–
16.00–17.00	17.00	17.00

Notes

- If the patient develops symptoms of water loss, or loses more than 3% of initial body weight, samples for serum and urine osmolality should be collected immediately.
- If serum osmolality exceeds 295 mmol/kg and/or serum sodium exceeds 145 mmol/L, the test should be discontinued and the vasopressin test performed.
- If the test runs to completion and the urine osmolality remains below 600 mmol/kg, the supplementary vasopressin test should be performed.

Vasopressin test

Give 2 µg of dDAVP i.v. The patient is allowed to drink, but total fluid intake should be restricted to 1000 mL until 09.00 h the next morning, unless the patient's weight loss continues above 3%, when free fluids should be allowed. Further urine collections are made the same day at 19.00 and 22.00 h, and on the following day at 07.00 and 09.00 h.

Interpretation

A normal subject will concentrate urine to >600 mmol/kg during the period of water deprivation, and the serum

osmolality will remain within the physiological range (or, more strictly, the urine:serum osmolality ratio will be greater than 2:1). If urine osmolality fails to increase to >600 mmol/kg, but increases following dDAVP by >20%, then cranial diabetes insipidus (CDI) is likely. The weight loss recorded should always equate with total urine volume passed.

Reference

This protocol is modified from the original description of the 'short' water deprivation test by Dashe AM, Cramm RE, Crist AC, Habener JF, Solomon DH. A water deprivation test for the differential diagnosis of polyuria. *J Am Med Assoc* 1963;185:699–703.

(b) Hypertonic saline infusion

This test is used to directly assess the osmoregulatory control of the release of plasma AVP and/or to assess the control of subjective thirst.

The patient is food-fasted overnight (12 h), but is allowed free access to water. Smoking is not allowed during this 12 h period or during any part of the test. No fluid of any kind should be consumed during the test, including sips, mouthwashes or ice cubes: all of these oropharyngeal stimuli can suppress the release of arginine vasopressin (AVP) from the posterior pituitary.

Pre-infusion preparation

09.00 h	Weigh Supine position Indwelling cannulae with both antecubital veins kept patent with heparinized saline	One for eventual infusion of hypertonic saline, one for blood sampling
09.00–10.00 h	Rest hour	

Infusion protocol

Time of blood collection (h)	Tests
10.00	AVP, osmolality, urea and electrolytes
10.30	"
11.00	"
11.30	"
12.00	"

Commence 5% saline (0.06 mL/kg/min) infusion into one cannula at 10.00 h, following the first blood sample.

For interpretation of results, see [Figure 4.3](#) and text.

Notes

- If information is required on subjective thirst rating, then the patient should be shown, at each time blood is sampled, a unitless 100 mm scale with a maximum limit labelled 'Extreme thirst' and a minimum limit 'No thirst'. The patient is asked to indicate thirst levels on the scale. Two separate scales should be

completed on each occasion to obtain a measure of precision.

- Information concerning the handling of blood samples for AVP should be obtained from the laboratory. In general, samples should be collected into pre-chilled heparin tubes, transported on ice to the laboratory immediately, centrifuged rapidly at 4°C and the plasma stored at a maximum temperature of -20°C (preferably -70°C). The time from collection to storage should not usually exceed 20 min.
- Patients with a history of congestive cardiac failure should be closely monitored and, if necessary, the test curtailed and furosemide (40 mg i.v.) administered.

Reference

Adapted from Robertson GL, Athar S. The interaction of blood osmolality and blood volume in regulating plasma vasopressin in man. *J Clin Endocrinol Metab* 1976;42:613–20.

(c) Water load test

This test is used to assess osmoregulation indirectly by determining the renal response to water loading.

The patient is allowed free access to fluid 12 h prior to the test to ensure adequate patient-determined hydration at commencement. No smoking is allowed during the test period or for the 12 h prior to this. Adequate glucocorticoid replacement is required for patients with hypoadrenalism.

At 09.00 h, the bladder is emptied and an aliquot of urine (10–15 mL) saved for osmolality measurement. Blood (for urea and electrolytes, and osmolality) is collected and the patient weighed. An oral water load (20 mL/kg) is then consumed within 20 min. Further samples and weighings are obtained according to the following schedule. Urine volume output is measured accurately each hour.

Sample	Time of urine collection (h)	Time of blood collection (h)	Time of weighing (h)
1	09.00	09.00	09.00
2	09.00–10.00	–	–
3	10.00–11.00	–	–
4	11.00–12.00	–	–
5	12.00–13.00	13.00	13.00

Interpretation

In a normally hydrated subject, over 80% of a standard water load is excreted within 4 h of ingestion. Urine osmolality will typically fall to <100 mmol/kg. Weight changes should confirm the observed difference between the water volume ingested and the urine volume loss. Difficulties in micturition may make the test difficult or impossible to interpret.

Reference

Penney MD, Murphy D, Walters G. Resetting of the osmoreceptor response as a cause of hyponatraemia in acute idiopathic polyneuritis. *Br Med J* 1979;2:1474–6.

Hydrogen ion homoeostasis and tissue oxygenation and their disorders

William J. Marshall

CHAPTER OUTLINE

INTRODUCTION 65	Respiratory acidosis 80
THE PHYSIOLOGICAL ROLE OF HYDROGEN IONS 65	Non-respiratory alkalosis 82
Definitions 65	Respiratory alkalosis 84
HYDROGEN ION HOMOEOSTASIS 66	The interpretation of acid–base data 85
Buffering 66	Mixed disorders of hydrogen ion homoeostasis 86
Hydrogen ion turnover 68	TISSUE OXYGENATION 87
Hydrogen ion production 68	Introduction 87
Hydrogen ion excretion 70	Pulmonary function 87
Summary 73	The role of haemoglobin in oxygen transport 88
THE ASSESSMENT OF ACID–BASE STATUS 73	The effects of pulmonary disease on oxygen uptake into blood 89
Clinical assessment 73	Oxygen transport to tissues 89
Laboratory assessment 73	Hypoxia 90
DISORDERS OF HYDROGEN ION HOMOEOSTASIS 74	CONCLUSION 92
Introduction 74	
Non-respiratory acidosis 74	

INTRODUCTION

Disorders of both hydrogen ion homoeostasis and tissue oxygenation will be discussed in this chapter. Abnormalities of hydrogen ion homoeostasis occur frequently in respiratory disorders, as a result of changes in the rate of excretion of carbon dioxide. Such disorders can also affect oxygenation, and impaired tissue oxygenation is an important potential cause of acidosis. Furthermore, hydrogen ion concentration, carbon dioxide and oxygen are measured using related technologies, usually with the same instrument.

The first part of this chapter deals with hydrogen ion homoeostasis and its disorders (colloquially often referred to as ‘acid–base balance’ and ‘acid–base disorders’, respectively), while the second part deals with the mechanism whereby oxygen is made available to the tissues, disorders in which tissue oxygenation is impaired and how tissue oxygenation is measured.

THE PHYSIOLOGICAL ROLE OF HYDROGEN IONS

Hydrogen ions are ubiquitous in the body and maintenance of appropriate concentrations is critical to normal function. The gradient of hydrogen ion concentration

between the inner and outer mitochondrial membrane drives oxidative phosphorylation; changes in hydrogen ion concentration can affect the surface charge and physical conformation of proteins and thus their function, and hydrogen ion concentration determines the degree of ionization of weak acids and bases and can thus affect the disposition of such substances, among which are many with important physiological and pharmacological functions.

The hydrogen ion concentration of the blood is normally controlled within narrow limits, in health rarely exceeding 46 nmol/L or falling below 35 nmol/L. This regulation is achieved in spite of the continuous production of hydrogen ions as a result of the normal processes of metabolism. Intracellular hydrogen ion concentration is, in general, higher; in the cytosol being slightly so while in lysosomes it is several orders of magnitude higher. However, the interior of mitochondria is slightly alkaline.

Definitions

An increase in the hydrogen ion concentration of the blood is termed acidaemia and a decrease, alkalaemia. The term ‘acidosis’ strictly describes a pathological disturbance that can result in acidaemia, but may not necessarily do so

because of the simultaneous existence of another disturbance (possibly the result of a physiological compensatory process) that has an opposing effect. Similarly, alkalaemia is not always present in alkalosis. These distinctions, although made much of by some authors, often only introduce confusion and will not be pursued in this chapter.

Strictly speaking, it is the *activity* of hydrogen ions and not their *concentration* that is relevant, and devices for measuring hydrogen ions respond to their activity. Activity and concentration are only the same in ideal solutions, which biological fluids are not, but the distinction can be ignored for practical purposes.

Hydrogen ion concentration can also be expressed in terms of pH. The pH of a solution is the logarithm (base 10) of the reciprocal of hydrogen ion concentration (or minus the logarithm of hydrogen ion concentration). Thus, a solution with hydrogen ion concentration of 100 nmol/L (100×10^{-9} mol/L) has a pH of 7.00 ($\log_{10} 1/100 \times 10^{-9}$). pH does not have units; it varies in a reciprocal and nonlinear fashion with hydrogen ion concentration. The reference range for the pH of blood, corresponding to the range for hydrogen ion concentration given above, is 7.36–7.42. It is necessary, in discussing hydrogen ion homeostasis and its disorders, to consider the production and disposal of hydrogen ions; it is therefore logical to discuss the effects of changes in these processes on hydrogen ion concentration rather than on a derived unit. Furthermore, as will be seen, the analysis of disorders of hydrogen ion homeostasis is facilitated considerably if direct measurements are used.

In health, the rates of formation of hydrogen ions and of their consumption and excretion are in balance, but disturbances of hydrogen ion homeostasis occur frequently and in a wide variety of disease states. They are classified traditionally as respiratory or non-respiratory in origin, according to whether the primary abnormality is the result of an excess or deficiency of carbon dioxide (respiratory) or of bicarbonate (non-respiratory). Non-respiratory disturbances are often referred to as metabolic. As will be seen, however, a respiratory disorder that causes hypoxia may produce an acid–base abnormality whose characteristics are non-respiratory, or even combine respiratory and non-respiratory features. Nevertheless, the distinction is a useful one and is of considerable value in analysing disorders of hydrogen ion homeostasis.

HYDROGEN ION HOMOEOSTASIS

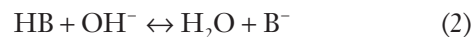
Buffering

What follows is a brief, simplified account of the essential features of buffering, sufficient for an appreciation of the physiology of hydrogen ion homeostasis.

In essence, buffer systems limit the extent to which the hydrogen ion concentration actually changes in the face of any tendency to change. A buffer system (or buffer pair) consists of a weak (that is, only partly dissociated) acid and its conjugate base (that is, the anion that combines with a hydrogen ion to form the acid). If the acid is HB and the conjugate base is B^- , the relevant reactions are:



and



Bicarbonate

An example of central importance to physiology is the carbonic acid–bicarbonate system. Carbonic acid is a weak acid, which partly dissociates to bicarbonate and hydrogen ion:



If an amount of hydrogen ions were to be added to a solution containing this buffer pair, some would combine with bicarbonate to form carbonic acid, thus limiting the increase in hydrogen ion concentration that otherwise would be expected to occur, but at the expense of consuming bicarbonate. Removal of hydrogen ions (or addition of hydroxyl, which, by combining with hydrogen ions to form water, has an identical effect) would cause dissociation of carbonic acid, producing hydrogen ions (and bicarbonate) and thus limiting the fall in hydrogen ion concentration that otherwise would be expected to occur.

It will be apparent that if a buffer is to function effectively in the face of equal tendencies for hydrogen ion concentration to increase and decrease, the concentrations of the acid and conjugate base should be equal, for example for the acid HB, $[HB] = [B^-]$. Since at equilibrium:

$$[H^+] = K_a [HB] / [B^-] \quad (4)$$

(where K_a is the equilibrium constant), it follows that a buffer will be most effective when the hydrogen ion concentration to be defended is numerically similar to the equilibrium constant. It should also be appreciated that hydrogen ion concentration, $[H^+]$, depends not on the absolute values for the concentrations of acid and conjugate base, but on the *ratio* of their concentrations. On the other hand, the buffer capacity, that is, the extent to which the buffer can absorb hydrogen ions, clearly will depend upon the absolute values of these concentrations.

For bicarbonate, it follows from the general equation (Eqn. 4) that:

$$[H^+] = K_a [H_2CO_3] / [HCO_3^-] \quad (5)$$

The concentration of carbonic acid cannot readily be determined, but it is directly proportional to the concentration of carbon dioxide $[CO_2]$, since carbonic acid is formed by the hydration of carbon dioxide:



Thus, Equation 5 can be rewritten in the form:

$$[H^+] = K_a' [CO_2] / [HCO_3^-] \quad (7)$$

where K_a' is a constant numerically incorporating the equilibrium constants for the reactions represented by Equations 3 and 5. The value of K_a' is approximately 800 nmol/L and, at physiological extracellular hydrogen ion concentration (40 nmol/L), the numerical value of

the molar ratio $[\text{CO}_2]/[\text{HCO}_3^-]$ is approximately 0.05, suggesting that this system should be a poor buffer, particularly against a tendency for hydrogen ion concentration to fall. However, the body is a net producer of acid, that is, the tendency is for hydrogen ion concentration to rise. More importantly, carbon dioxide, generated by the buffering of hydrogen ion by bicarbonate and the subsequent dissociation of carbonic acid, can be removed by the lungs, keeping the carbon dioxide concentration constant. As a result, the effective buffering capacity of the carbonic acid–bicarbonate system is greatly increased, and it is a vitally important physiological buffer, particularly in the extracellular fluid.

In practice, carbon dioxide *concentration* cannot be readily determined, but it is related to the partial pressure, PCO_2 , such that $[\text{CO}_2] = 0.225 \times PCO_2$ when PCO_2 is measured in kilopascals (kPa).

Phosphate

The monohydrogen and dihydrogen phosphate ions (HPO_4^{2-} and H_2PO_4^-) form a buffer pair with K_a being approximately 160 nmol/L. Although the K_a appears relatively favourable for buffering at physiological hydrogen ion concentrations, the concentration of phosphate in the extracellular fluid is too low for it to be of significance in this respect. Phosphate is, however, an important buffer in the urine, where its concentration is much greater.

Haemoglobin

All proteins can buffer hydrogen ions to some extent by virtue of their content of polar amino acid residues. Haemoglobin (Hb) is an important buffer. The reaction can be represented as:



although each haemoglobin molecule is capable of buffering a number of hydrogen ions. It is relevant to note that haemoglobin is a more effective buffer for hydrogen ions when it is in the deoxygenated rather than the oxygenated form, and that oxygen release is facilitated by the buffering (the Bohr effect).

The efficacy of haemoglobin as a buffer is enhanced by the presence in erythrocytes of the enzyme carbonate dehydratase, which catalyses the hydration of carbon dioxide (Eqn. 6). Despite the fact that they are responsible for the majority of oxygen transport in the blood, erythrocytes obtain their energy anaerobically, by glycolysis, and thus do not generate carbon dioxide. In the capillary beds of tissues, carbon dioxide produced by aerobic metabolism readily diffuses down the concentration gradient into erythrocytes, where it is hydrated to form carbonic acid, which dissociates to form bicarbonate and hydrogen ions (Fig. 5.1). The latter are buffered by haemoglobin while bicarbonate diffuses out of the cells in exchange for chloride, so that the products of the dissociation of carbonic acid are removed, allowing further dissociation to occur and thus, by a mass action effect, stimulating its formation. In the lungs, alveolar PCO_2 is lower than venous PCO_2 and

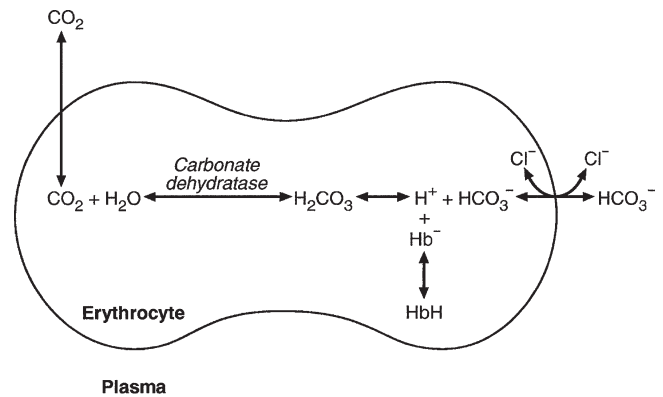


FIGURE 5.1 ■ Transport and buffering of carbon dioxide in erythrocytes. Carbon dioxide is converted into carbonic acid in erythrocytes; this dissociates to form bicarbonate, which diffuses into the plasma in exchange for chloride and hydrogen ions, which are buffered by haemoglobin. In the alveoli, the reverse process liberates carbon dioxide.

the process reverses, carbon dioxide being generated and excreted, while the release of hydrogen ions from haemoglobin favours oxygen uptake.

Thus, in addition to transporting carbon dioxide, the conversion of carbon dioxide to bicarbonate serves to minimize the potential change in the value of the ratio $[\text{CO}_2]/[\text{HCO}_3^-]$ (and hence in hydrogen ion concentration) between arterial and venous blood. However, although erythrocytes are a potential source of bicarbonate, they can only generate bicarbonate if the hydrogen ion produced simultaneously can be buffered. There is clearly a limit to the extent to which this can occur, imposed by the buffering capacity of haemoglobin, so that this mechanism can have only a limited role in the correction of an acidosis. The way in which such correction is achieved is considered in a later section.

Other proteins

Although plasma proteins buffer hydrogen ions, their molar concentrations are lower than that of haemoglobin and their buffering capacity is less. In contrast, intracellular and tissue proteins make an important contribution to buffering. In chronic acidosis, they may contribute up to one-third of total buffer capacity. Buffering by bone is particularly important.

Ammonia

It is often stated that ammonia is an important buffer in the urine because it can combine with hydrogen ions to form ammonium ions. This cannot be true. Ammonium is a very weak acid, with an equilibrium constant approximately 100 times lower than physiological hydrogen ion concentration, so that when ammonia is produced in the body, it is immediately and almost completely converted to ammonium ions. The urinary excretion of ammonium is of relevance to hydrogen ion homeostasis because it represents a route for the disposal of ammonium which, unlike urea synthesis, does not result in the generation of hydrogen ions (see p. 70).

Hydrogen ion turnover

The fact that, on a normal diet, the daily excretion of hydrogen ions by the kidneys (the only physiologically important route of excretion) is 40–80 mmol tends to distract from the fact that there is a massive endogenous turnover of hydrogen ions. In the resting adult, intermediary metabolism accounts for a hydrogen ion turnover of 2500–3000 mmol/24 h (Table 5.1).

Normally, the rates of hydrogen ion formation and utilization during intermediary metabolism are in overall balance, although any discrepancy can have a major effect on hydrogen ion concentration. But even this turnover of hydrogen ions appears insignificant in comparison with that which is associated with the turnover of adenine nucleotides – the movement of hydrogen ions that takes place across the mitochondrial membrane during oxidative phosphorylation and the synthesis and hydrolysis of ATP. This has been estimated at 500 mol/24 h. The potential for a disturbance in these processes to cause an acidosis is clearly colossal, although, in health, the rates of adenine nucleotide reduction and oxidation and of ATP formation and utilization are equal, so that these processes have no net effect on hydrogen ion homeostasis. This may not, however, be true in disease.

As a result of oxidative metabolism, carbon dioxide is produced and excreted by the lungs. The rates of formation and excretion are normally equal, but carbon dioxide can combine with water to form carbonic acid, and the daily production of carbon dioxide in a resting adult is a potential source of approximately 15–20 mol of hydrogen ions. As has been alluded to, disorders affecting the excretion of carbon dioxide are an important cause of abnormalities of hydrogen ion homeostasis.

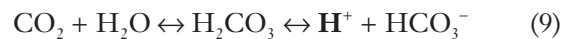
Tendencies for hydrogen ion concentration to change can be limited to some extent by buffering, but this process can only offer a temporary solution to an imbalance between the rates of hydrogen ion production and disposal, because the body's buffers have a limited capacity. Physiological processes often bring about a partial reversal of a change in hydrogen ion concentration (see Compensation, below), but ultimate correction of any disturbance requires equalization of the rates of production and disposal of hydrogen ion.

Hydrogen ion production

The processes involved in hydrogen ion production and utilization are summarized in Table 5.1. They comprise: processes involving carbon dioxide formation, reactions of intermediary metabolism and processes involving 'fixed' acids. In addition, and responsible for the bulk of hydrogen ion turnover, are the reactions involved in the complete oxidation of energy substrates. These processes are interlinked, but it is instructive to consider them separately.

Carbon dioxide

The role of carbon dioxide in relation to the formation of hydrogen ions has been mentioned above. Carbon dioxide, produced by oxidative metabolism, can become hydrated to carbonic acid, a weak acid that partly dissociates to hydrogen ion and bicarbonate:



The equilibrium for this reaction strongly favours carbon dioxide and water, but in tissues containing carbonate dehydratase (e.g. tubular cells in the kidneys), the rate of formation of carbonic acid is increased and it can become an important source of bicarbonate and hydrogen ions.

Incomplete metabolism of glucose: glycolysis and lactate metabolism

The most familiar process of intermediary metabolism that results in the formation of hydrogen ions is anaerobic glycolysis, the metabolism of glucose to lactate. The overall equation for this reaction is:



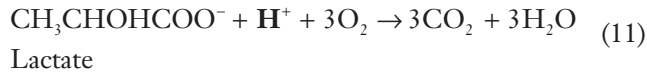
This process, which takes place particularly in skeletal muscle and erythrocytes, results in the formation of ~1.3 mol of hydrogen ions per 24 h in a 70 kg man at rest. The major route of disposal of lactate is glucose synthesis by gluconeogenesis in the liver and kidneys. The overall equation for this process is the reverse of that for glycolysis (most enzymes involved being common to both

TABLE 5.1 Approximate daily production and elimination of acid in a resting adult

Type of acid	Production	mmol/24 h	Disposal	mmol/24 h
Carbon dioxide	Tissue respiration	~17 500	Excretion by lungs	~17 500
Lactate	Glycolysis	~1300	Gluconeogenesis, oxidation	~1300
Free fatty acids	Lipolysis	~650	Re-esterification, oxidation	~650
Ketoacids	Ketogenesis	~500	Oxidation	~500
H ⁺ generated during urea synthesis	Ureagenesis	~1100	Oxidation of amino acids	~1100
Sulphuric and phosphoric acids	Metabolism of sulphur- and phosphate-containing amino acids	~40	Renal excretion ('fixed' or buffered acid)	~40

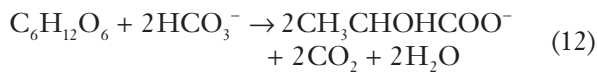
Note that all figures are approximate, being dependent on diet, body mass index, physical activity etc.

pathways, although some are unique). Gluconeogenesis thus consumes hydrogen ions. In health, lactate production and disposal are equal and so, too, are the production and disposal of hydrogen ions by these pathways. At rest, most of the lactate produced is converted back to glucose; during exercise, when lactate production is increased, 50% or more is completely oxidized instead. This process also consumes the hydrogen ions that are generated in its production and results in the formation of carbon dioxide and water:

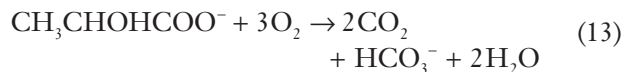


Thus, when lactate production and utilization are in balance, there is no net production of hydrogen ions. However, in disease states, an imbalance between these processes can be responsible for the development of acidosis.

The hydrogen ions generated in lactate production will be buffered principally by bicarbonate, so that Equation 10 could be written:



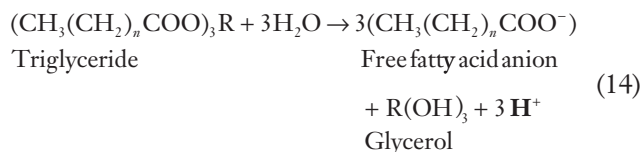
Clearly, the reverse reaction, gluconeogenesis, will regenerate the bicarbonate, and Equation 11 can be rewritten to show that this also occurs with complete oxidation of lactate:



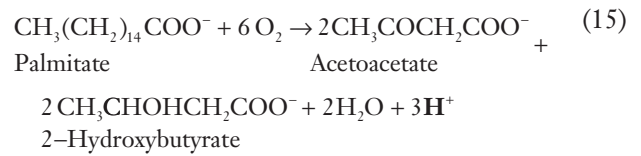
It is pertinent to point out that hyperlactataemia does not automatically indicate an acidosis. If it is a result of increased lactate production or decreased excretion, acidosis will only be present if the hydrogen ions produced simultaneously are not removed by excretion or metabolism. Intravenous fluids containing sodium lactate (e.g. Hartmann's solution) are actually a source of alkali, because the lactate ions are metabolized to bicarbonate (Eqn. 13).

Incomplete metabolism of triglycerides: ketogenesis

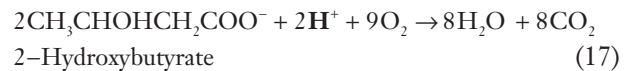
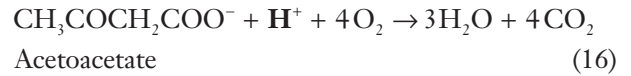
The liberation of free fatty acids from triglyceride (triacylglycerol) in adipose tissue results in the generation of hydrogen ions. The process is exemplified by the equation:



In the liver and adipose tissue, free fatty acids can be re-esterified to triacylglycerol, a process that consumes three hydrogen ions for each molecule of triacylglycerol synthesized. The further metabolism of free fatty acids to ketones in the liver (ketogenesis) also results in hydrogen ion production. An example of the reaction, starting from palmitate, is:



In health, this process may account for up to 0.4 mol hydrogen ion per 24 h, although in pathological states, its contribution may be much greater. However, ketoacids (strictly, oxoacids) are utilized as energy sources by skeletal muscle and other tissues. Their oxidation consumes hydrogen ions so that, in health, the overall rates of hydrogen ion production and disposal are equal:



In disease, notably in diabetic ketoacidosis, excessive ketogenesis is an important cause of acidosis. Ketonuria may exacerbate the acidosis: ketoacid anions are a potential source of bicarbonate so that their loss in urine (in maximally acidic urine, about half the ketoacids are present in this form) effectively reduces the potential for bicarbonate generation.

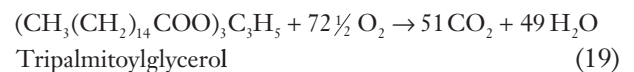
Complete oxidation of glucose and triglycerides

Glucose can also be completely oxidized to carbon dioxide and water; indeed, this is the major route of glucose metabolism in the body.



This is a complex process: oxidation is indirect, involving the transfer of hydrogen ions to adenine nucleotides, which are mainly oxidized by the mitochondrial electron transport system. Electrons are transferred to oxygen and combination with hydrogen ions produces water, while the energy released is transferred to ATP. But as Equation 18 indicates, the complete oxidation of glucose to carbon dioxide does not give rise to net formation of hydrogen ions: the products of its oxidation are carbon dioxide and water.

The same is true for the complete oxidation of triacylglycerols:



As with glucose oxidation, the process is far more complicated than appears from Equation 19, again involving the reduction of nucleotides, transfer of electrons to molecular oxygen, formation of water and trapping of energy in ATP, but the net production of hydrogen ions is zero.

Amino acid metabolism

Amino acid metabolism both produces and consumes hydrogen ions, according to the type of amino acid concerned. The metabolism of neutral amino acids

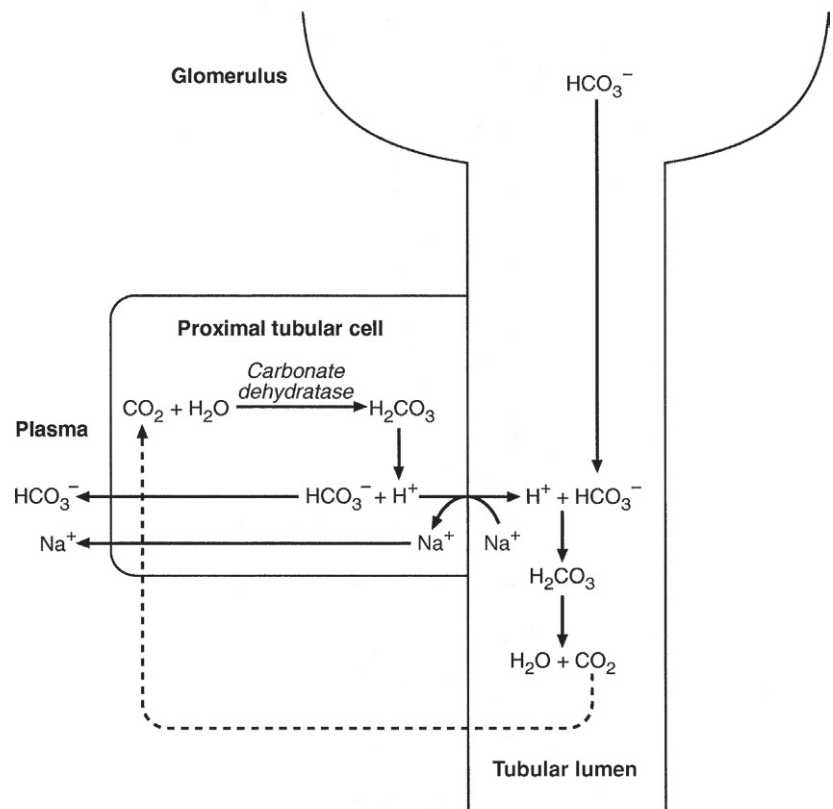
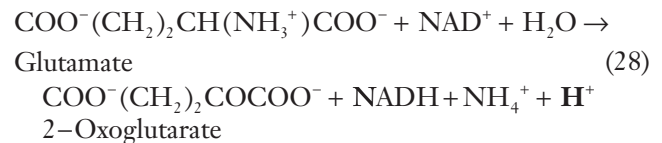
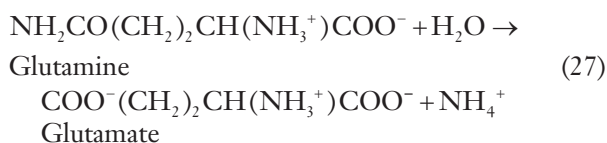


FIGURE 5.2 ■ Reabsorption of filtered bicarbonate. Bicarbonate ions cannot be directly reabsorbed. Instead, hydrogen ions and bicarbonate are generated from carbon dioxide in tubular cells; the hydrogen ions are secreted into the tubular lumen in exchange for sodium and titrate the filtered bicarbonate. Some of the carbon dioxide thus formed diffuses into renal tubular cells. The bicarbonate formed in the cells accompanies sodium as it is pumped into the interstitial fluid and thence reaches the plasma.

of approximately 4.5, equivalent to a hydrogen ion concentration of $38\ \mu\text{mol/L}$. This represents a 1000-fold concentration gradient with respect to the extracellular fluid, but clearly, excretion of free hydrogen ions alone would be insufficient to remove the daily burden of acid produced by metabolic processes, which is measured in millimoles. Significant acid excretion is achieved by hydrogen ions being buffered by phosphate, titrating monohydrogen phosphate (HPO_4^{2-}) (the principal form in the plasma and thus the glomerular filtrate) to dihydrogen phosphate (H_2PO_4^-) ions. It should be noted that, since the formation of hydrogen ions in renal tubular cells is accompanied by stoichiometric generation of bicarbonate ions, the excretion of hydrogen ions additionally results in the regeneration of bicarbonate ions and thus restores buffering capacity. This process is summarized in [Figure 5.3](#).

The role of urinary ammonium excretion. Although the urine contains ammonium and, indeed, the amount excreted increases considerably in states of chronic acidosis, this cannot, for the reasons indicated above, in itself, constitute net hydrogen ion excretion. Ammonium is produced in renal tubular cells by the action of the enzyme glutaminase on glutamine, the amide of glutamic acid ([Eqn. 27](#)), and the oxidative deamination of glutamate by glutamate dehydrogenase ([Eqn. 28](#)).



Glutamate is formed by transamination of 2-oxoglutarate with other amino acids, a process that does not involve either the production or utilization of hydrogen ions, and the equation for glutamine synthesis is the reverse of [Equation 27](#). Thus, glutamine synthesis is a mechanism for the disposal of ammonium ions that does not (unlike urea synthesis) produce hydrogen ions. Although subsequent urinary ammonium excretion appears to be a means whereby hydrogen ions can be excreted in a buffered form, it does not represent direct excretion of hydrogen ions: rather, it is a process through which nitrogen can be excreted without the concomitant generation of hydrogen ions. As indicated in [Equations 27 and 28](#), the production of ammonium from glutamine also yields 2-oxoglutarate: this is a substrate for gluconeogenesis, a process that consumes hydrogen ions.

In acidosis, hepatic glutamine synthesis is increased: in the kidneys, the formation of ammonium from glutamine, urinary ammonium excretion and gluconeogenesis are all increased. The net result is a decrease in hydrogen ion formation and an increase in bicarbonate generation, both of which tend to correct the acidosis. Renal ammonium excretion is illustrated in [Figure 5.4](#).

The luminal membranes of renal tubular cells are actually impermeable to ammonium ions, but are permeable to ammonia. Continued diffusion of small amounts

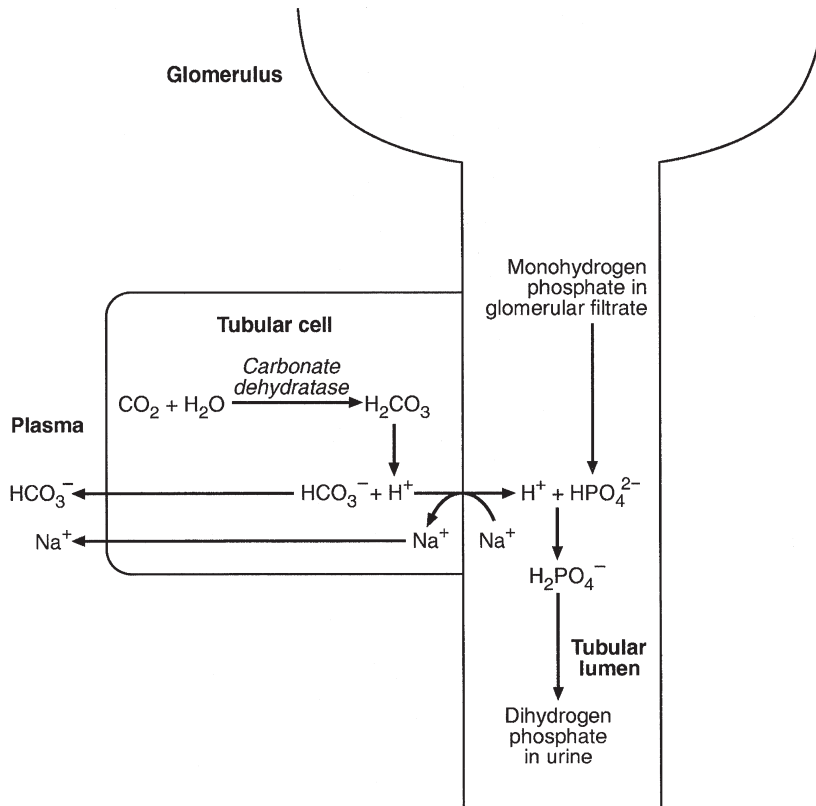


FIGURE 5.3 ■ Renal hydrogen ion excretion. Carbonic acid is generated from carbon dioxide and water and dissociates into hydrogen ions and bicarbonate. The hydrogen ions are secreted into the tubular lumen in exchange for sodium, where they titrate monohydrogen phosphate to dihydrogen phosphate. Bicarbonate and sodium ions are pumped from the cells into the interstitial fluid and thence reach the plasma.

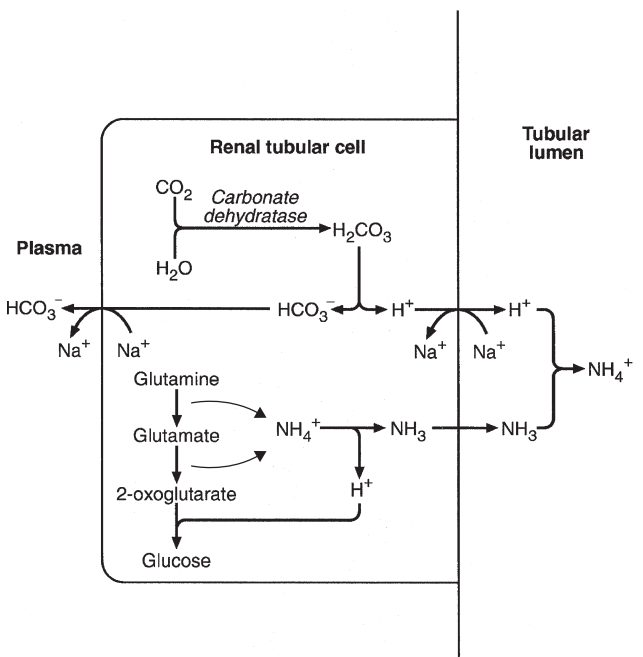


FIGURE 5.4 ■ Renal ammonium excretion. Ammonium is formed directly from glutamine and glutamate and is excreted in the urine. There is no net excretion of hydrogen ion. This process is accompanied by the formation of oxoglutarate, which can undergo gluconeogenesis, and results in the consumption of hydrogen ions that are produced with bicarbonate from carbonic acid.

of ammonia in equilibrium with ammonium within renal tubular cells into the tubular fluid, results in continued formation of ammonia from ammonium. In the lumens of nephrons, this ammonia is immediately converted back to ammonium ions. This process does not entail *net* excretion of hydrogen ions; indeed, the *net* result is the same as if ammonium ions *were* transported directly.

The role of the liver in hydrogen ion homeostasis. Traditionally, the kidneys have been considered (together with the lungs) as the major organs responsible for hydrogen ion homeostasis, but the liver also plays a role, although its extent remains controversial. While the kidneys are the only organs capable of excreting hydrogen ions from the body, the liver both generates and consumes hydrogen ions. As indicated above, it has a central role in the production (e.g. ketogenesis, ureagenesis) and utilization of hydrogen ions (e.g. gluconeogenesis), yet an abnormal hydrogen ion concentration is an uncommon finding in patients with liver failure unless this is severe or there is accompanying renal disease. But this should not be surprising, since it is well known that the liver has substantial reserve capacity, and functional impairment may only occur with massive liver damage. There is certainly evidence that acid-base status has an influence on hepatic urea synthesis and glutamine synthesis, and affects lactate and ketone metabolism. For example, acidosis tends to stimulate hepatic glutamine synthesis and lactate disposal, but inhibits ketogenesis.

Summary

Three organs are involved in hydrogen ion homeostasis: the lungs, the kidneys and the liver. Extracellular hydrogen ion concentration depends on the ratio $PCO_2 / [HCO_3^-]$. The respiratory centre is exquisitely sensitive to arterial PCO_2 and, given normal respiratory function, PCO_2 is maintained within narrow limits by changes in the rate and depth of respiration. It has been calculated that cessation of carbon dioxide excretion would lead to potentially fatal acidosis within 30 min (although this is purely hypothetical, since the accompanying failure in oxygen supply would be fatal before this time).

Nearly all the carbon dioxide produced each day is excreted through the lungs in the expired gas, although the liver is capable of disposing of a very small amount through the anaplerotic carboxylation of pyruvate to oxaloacetate. Although the liver can both generate and utilize hydrogen ions in metabolism, only the kidneys can excrete hydrogen ions from the body. In terms of the total turnover of hydrogen ions, the amount excreted this way is small. Nevertheless, it is vitally important, and a failure of renal hydrogen ion excretion frequently leads to acidosis, although even in acute kidney injury, if there are no additional causes of acidosis, it can take several days before the acidosis itself becomes severe.

Whilst any one of abnormal respiratory, hepatic or renal function may lead to the development of acidosis or alkalosis, changes in the others may ameliorate the effect on arterial hydrogen ion concentration. The compensatory hyperventilation seen in patients with non-respiratory acidosis is an example of this. Such compensatory processes are of vital importance in disturbances of hydrogen ion homeostasis, and are discussed in relation to the specific conditions in which they occur. However, it is important to appreciate that, although they may restore an abnormal hydrogen ion concentration to, or at least towards, normal, they do not correct the underlying disturbance.

THE ASSESSMENT OF ACID–BASE STATUS

Clinical assessment

None of the clinical features of acidosis and alkalosis is specific to these disturbances and they may only be present when the disturbances are severe. The conditions that give rise to acidosis and alkalosis may have specific features, but, while clinical assessment is important in patients with disturbances of hydrogen ion homeostasis, laboratory investigations are vital for their diagnosis, the assessment of their severity and for monitoring their progress.

Laboratory assessment

Hydrogen ion concentration and PCO_2

Measurements of arterial blood hydrogen ion concentration (strictly, activity) and $PaCO_2$ (the 'a' indicates arterial, and 'A' alveolar; in this chapter, for the sake of simplicity, the 'a' is omitted except where doing so could lead to ambiguity) are fundamental to the assessment of acid–base

status. Analysers may express hydrogen ion concentration as pH, but the SI unit is concentration and the use of concentration greatly facilitates data interpretation.

Since the hydrogen ion concentration is determined by the ratio of PCO_2 to bicarbonate concentration, bicarbonate concentration is not an independent variable and knowledge of it is not necessary for the characterization of acid–base disorders. Analysers calculate bicarbonate from PCO_2 , $[H^+]$ and K_a' , using the formula given in Equation 7. However, although K_a' is supposedly a constant, there is evidence that it can vary unpredictably, particularly in severely ill patients.

True bicarbonate concentration is not readily measurable. Most laboratory measurements of 'bicarbonate' are in fact measurements of total carbon dioxide (TCO_2), to which bicarbonate makes by far the greatest contribution, but which also measures dissolved carbon dioxide, carbonic acid and carbamino compounds (histidine residues in proteins that have combined with carbon dioxide). Together, all these other species contribute only about 10% to the TCO_2 . It should be noted that the TCO_2 concentration of plasma and serum decreases in vitro, as carbon dioxide is lost to the atmosphere. Caution should thus be exercised in interpreting its value if there has been any significant delay between the times of the blood sample being taken and the plasma or serum being analysed. Arterial blood for analysis must be suitably anticoagulated; any gas bubbles must be expelled and the specimen protected from the atmosphere and either analysed immediately using a point of care instrument, or transported rapidly to the laboratory, chilled by placing the syringe in which it is collected in iced water.

Derived variables

Analysers that measure hydrogen ion concentration and PCO_2 are generically termed 'blood gas analysers'; they also measure PO_2 and may provide other information of value in determining arterial oxygen content. They also frequently generate various derived terms, including 'standard bicarbonate', 'base excess' and 'standard base excess'. These are calculated terms and add nothing to the characterization of disorders of hydrogen ion homeostasis beyond what can be determined from consideration of $[H^+]$ and PCO_2 .

As will be seen, bicarbonate can be affected by both respiratory and non-respiratory disturbances. The standard bicarbonate is the bicarbonate concentration that would be expected in that blood sample were the PCO_2 to be normal. It supposedly eliminates any contribution by a respiratory disturbance, so that an abnormal standard bicarbonate concentration is taken as indicating the presence of a non-respiratory disturbance of hydrogen ion homeostasis. The base excess is the calculated amount of acid in millimoles that would have to be added to 1 L of the patient's blood in vitro to restore the hydrogen ion concentration to normal in an alkalosis (the base deficit is the corresponding figure for alkali in an acidosis). The standard base excess extends the base excess to include the whole extracellular compartment. The calculation of these derived variables makes unwarranted assumptions, and the data themselves can confuse, rather

than facilitate, the analysis of clinical disorders. The only *measurement* required to derive these variables in addition to PCO_2 and $[H^+]$ is the haemoglobin concentration. Since, therefore, derived units do not incorporate any other independent measurements of acid–base status, it follows that they cannot provide any information that cannot be derived from the measured variables alone. Furthermore, although it is correct that abnormal values for standard bicarbonate and base excess indicate the presence of a non-respiratory component in an acid–base disorder, *they do not distinguish between this being a primary disorder or a compensatory response to a primary respiratory disorder.*

Base deficit is sometimes used to calculate the amount of bicarbonate that is required to correct an acidosis, but, in practice, if an acidosis is so severe that bicarbonate treatment is warranted, this should be given in small amounts and the effect assessed by frequent measurements of $[H^+]$ and PCO_2 .

Anion gap

The anion gap is the sum of the concentrations (in mmol/L) of the two major cations in plasma minus those of the two major anions (i.e. $([Na^+] + [K^+]) - ([Cl^-] + [HCO_3^-])$). The total concentration of anionic charge in plasma must equal the concentration of cationic charge: the anion gap reflects mainly the anionic nature of most proteins in plasma at physiological hydrogen ion concentration, although phosphate and other anions make a small contribution. Its normal value is related to the concentrations of the species that determine it, and is ~15 mmol/L.

Determination of the anion gap can be of help in determining the cause of a non-respiratory acidosis. When acidosis is due to loss of bicarbonate (e.g. renal tubular acidoses, loss of bicarbonate from the gut), there is increased renal chloride retention; the lost bicarbonate is replaced by chloride and the anion gap is unaffected. On the other hand, when acidosis is due to ingestion or excess generation of acids, the associated anions (e.g. lactate) replace bicarbonate as this is consumed by buffering, and the anion gap is increased. In practice, determination of the anion gap is only likely to be helpful if the cause of the acidosis is not already obvious. It certainly does not substitute for careful clinical assessment. For its calculation, plasma chloride concentration must be known, and in the UK, this often must be requested individually, rather than being available as part of a profile of tests. Also, the imprecision of the value of the anion gap, summing as it does the individual imprecisions of the measurements of the analytes that are required to calculate it, needs to be taken into account.

Other investigations

Other biochemical measurements may be of value in the assessment of disorders of hydrogen ion homeostasis under certain conditions, for example in patients with diabetes, in neonates and in poisoned patients. These measurements are discussed in the sections that follow.

DISORDERS OF HYDROGEN ION HOMOEOSTASIS

Introduction

Although they differ in their pathogenesis, an understanding of the disorders of hydrogen ion homeostasis is facilitated by considering them in a similar way. The processes involved are the generation of the disorder, buffering, physiological compensation and ultimate correction. In practice, there is often overlap between these.

Disorders of hydrogen ion homeostasis may be simple, that is, involving only one type of disturbance, or mixed (two or more disturbances arising together). Unless rapidly corrected, simple disturbances typically generate secondary changes, which, in terms of bicarbonate concentrations and PCO_2 , can take on the characteristics of mixed disturbances.

Non-respiratory acidosis

This disorder can develop as a consequence of any, or a combination of, an increase in the rate of generation of hydrogen ions, a decrease in the rate of their utilization or excretion or a decrease in buffering capacity.

A list of some of the causes of non-respiratory acidosis is given in [Box 5.1](#). Frequently, more than one cause may contribute to an acidosis in an individual patient. For example, patients with diabetic ketoacidosis are usually hypovolaemic and their renal function is impaired. Many of these conditions are discussed in detail in other chapters of this book; the reasons why they can cause acidosis should be evident from the first section of this chapter. Some conditions of particular interest are discussed later. Although many of the conditions causing acidosis have distinct clinical features, some of the features of acidosis, and of the body's response to it, are common to them all. Non-respiratory acidosis can develop rapidly, particularly when it is due to increased production of hydrogen ions or loss of bicarbonate.

Compensatory responses in non-respiratory acidosis

Buffering. A tendency for hydrogen ion concentration to increase will be resisted by buffering; bicarbonate will be consumed and its concentration in the plasma will tend to fall. In the early stages of a condition that can cause acidosis, buffering, together with increased renal hydrogen ion excretion and bicarbonate regeneration, may prevent the development of a significant rise in hydrogen ion concentration. In chronic acidosis, buffering of hydrogen ions by tissue proteins also plays an important part in limiting the rise in hydrogen ion concentration.

Hyperventilation. Provided that respiratory function and its control are normal, acidosis stimulates ventilation through effects on aortic and carotid body chemoreceptors, and directly on the respiratory centre. Kussmaul breathing – deep, sighing respiration – is characteristic of non-respiratory acidosis. This not only allows excretion of carbon dioxide derived from the carbonic acid produced

BOX 5.1 Some causes of non-respiratory acidosis

Predominant increased acid formation

- Ketoacidosis^a
 - Diabetic
 - Alcoholic
 - Starvation
- Lactic acidosis^a
 - Hypoxic (type A)
 - Other causes (type B, see Box 5.2)
- Poisoning^a
 - Salicylate
 - Alcohols
- Inherited organic acidoses^a

Predominant decreased acid excretion

- Uraemic acidosis^a
- Renal tubular acidoses types 1 and 4 (see Table 5.3)

Predominant gain of acid or substances causing acidosis

- Ingestion of strong acid
- Ingestion/infusion of ammonium chloride
- Intravenous feeding with excess cationic amino acids
- Dilutional acidosis

Predominant loss of base

- Gastrointestinal loss
 - Diarrhoea
 - Pancreatic fistula
- Renal loss
 - Renal tubular acidosis type 2
 - Acetazolamide treatment
 - Ureteroenterostomy

^aAcidoses associated with a high anion gap. Note: In many instances, more than one mechanism may contribute to the acidosis.

by buffering, but actually reduces the PCO_2 , which in turn tends to lower the hydrogen ion concentration towards normal. The muscular effort of hyperventilation itself generates carbon dioxide, and there is a lower limit to the PCO_2 that can be obtained as a result of hyperventilation (approximately 1.4–1.6 kPa). Many causes of non-respiratory acidosis are progressive conditions, and the acidosis may be so severe that respiratory compensation cannot normalize the hydrogen ion concentration. However, when the acidosis is only mild, a new steady state may be achieved in which hyperventilation maintains a hydrogen ion concentration that is only slightly higher than normal.

The increased ventilation, and thus respiratory compensation for non-respiratory acidosis, develops rapidly, so that the predicted changes of a 'pure', that is uncompensated, non-respiratory acidosis (elevated blood hydrogen ion concentration, decreased bicarbonate and normal PCO_2) do not occur. However, compensation may take several hours to become maximal. This is because carbon dioxide equilibrates more rapidly across the blood–brain barrier than bicarbonate. Initial hyperventilation lowers cerebrospinal fluid (CSF) PCO_2 , and thus hydrogen ion concentration, and tends to counter the stimulation of respiration by peripheral chemoreceptors. Only as CSF bicarbonate concentration falls will its hydrogen

ion concentration rise and augment the respiratory drive. The reverse phenomenon may be seen if rapid correction of an acidosis is attempted.

In patients with respiratory impairment, the efficacy of the compensatory process may be greatly reduced.

Renal hydrogen ion excretion. Provided that renal dysfunction is not the cause of acidosis and renal function is not compromised, the urine hydrogen ion concentration rises to its maximum possible value. However, the capacity of the kidneys to excrete acid is limited: the fixed buffering capacity of the urine is approximately 2–3 times the normal acid load. In this process, filtered mono-hydrogen phosphate is titrated virtually completely to dihydrogen phosphate. In mild or transient acidosis, this may be sufficient to prevent a significant increase in blood hydrogen ion concentration occurring. If acidosis persists, however, there is an adaptive increase in the excretion of ammonium ions. Ammonium excretion may increase five-fold or more. As has been discussed above, although it has been widely considered that this represents the excretion of buffered hydrogen ions, it is more accurate to consider renal ammonium excretion as representing an alternative pathway for ammonium disposal that, unlike urea synthesis, does not involve the *generation* of hydrogen ions. Acidosis induces increased synthesis of glutaminase, the enzyme responsible for the formation of ammonium from glutamine. It is noteworthy in this respect that renal gluconeogenesis, which provides a pathway for the utilization of the 2-oxoglutarate derived from the carbon skeleton of glutamine, is also increased in acidosis.

Biochemical characteristics of non-respiratory acidosis

The cardinal features of a non-respiratory acidosis are an elevated blood hydrogen ion concentration and a decrease in bicarbonate. Hyperventilation results in a decrease in PCO_2 . The extent of the change in PCO_2 predicted for a given decrease in bicarbonate is discussed later. Comparison of the observed change with the predicted change can indicate whether there is an additional component to the acid–base disturbance.

In addition to these changes, other consequences of acidosis, especially hyperkalaemia, may be present, together with any specific features of the cause of the acidosis.

Systemic effects of acidosis

Many of the systemic effects of acidosis are common to acidosis of whatever cause. The effects on the cardiovascular system, oxygen delivery to tissues, the nervous system, potassium homeostasis and bone are of particular significance. Acidosis also has important effects on intermediary metabolism.

The cardiovascular system. Acidosis has a negative inotropic effect, although this is probably only of significance in severe acidosis. Acidosis causes arteriolar vasodilatation and constriction of peripheral veins, but in many forms of acidosis these responses are obscured by other influences on vasomotor tone.

Oxygen delivery to tissues. Acidosis causes a right shift in the oxyhaemoglobin dissociation curve (the Bohr effect, see p. 90) and this facilitates oxygen delivery to tissues. An increase in hydrogen ion concentration decreases erythrocyte 2,3-diphosphoglycerate (2,3-DPG) concentrations through effects on both synthesis and breakdown; this causes a left shift in the curve, but, whereas the Bohr effect is immediate, the fall in 2,3-DPG takes place over a matter of hours. The reverse is also true; so, if hydrogen ion concentration is restored to normal rapidly, oxygen delivery will be compromised until 2,3-DPG concentrations are restored to normal. This is a potential hazard if an attempt is made to correct an acidosis rapidly by the intravenous infusion of bicarbonate.

The nervous system. Patients with acidosis can demonstrate impaired consciousness of varying degrees of severity, but there is little correlation between this and the severity of the acidosis. In many patients with acidosis, other factors are operating that may affect CNS function, and changes in blood flow and oxygen delivery secondary to the acidosis may also be relevant.

Potassium homeostasis. There is a well-known association between acidosis and hyperkalaemia. This is multifactorial: movement of potassium ions from the intracellular to the extracellular compartment used to be thought to be related to intracellular buffering of hydrogen ions, but is probably to a greater extent the result of a loss of intracellular potassium to the plasma for other reasons, including a decrease in ATPase activity. It is of practical importance to note that total body potassium stores are frequently depleted, despite the high plasma concentration, and treatment of the acidosis may cause hypokalaemia if potassium is not replaced.

Bone. In chronic non-respiratory acidosis, there is significant buffering of hydrogen ions by bone, accompanied by decalcification, leading to a negative calcium balance. This is one factor contributing to renal osteodystrophy, the bone disease of chronic kidney disease. In addition, acidosis tends to increase ionized calcium concentration and so the filtered load presented to the renal tubules; renal calcium reabsorption may be decreased and calcitriol synthesis decreased.

Other effects. Acute acidosis can cause a leukocytosis; chronic acidosis can have a detrimental effect on nitrogen balance and insulin resistance can be a feature of both acute and chronic acidosis.

Management of non-respiratory acidosis

The logical management of a non-respiratory acidosis is to treat the underlying cause(s) of the disturbance appropriately. However, this may not always be possible and then, if the acidosis itself is having a significantly adverse effect on the patient, it may be necessary to attempt to lower the blood hydrogen ion concentration by giving alkali. The potential advantages are the reversal of the effects discussed above, including the hyperventilation, which may be particularly distressing for a conscious patient. There are, however, significant disadvantages. As

indicated above, rapid correction of an acidosis may have an adverse effect on oxygen delivery to tissues. When acidosis is associated with the presence of organic anions (e.g. lactate, acetoacetate), continued metabolism of these after correction of the acidosis (rather than *pari passu* with the correction, as would happen if the underlying cause were reversed) may, by consuming hydrogen ions, cause the blood hydrogen ion concentration to fall. This is a rebound phenomenon and not just a direct consequence of excessive alkali administration.

The alkali used most frequently when it is considered necessary to correct an acidosis is sodium bicarbonate. Buffering of hydrogen ions will result in the formation of carbon dioxide and blood PCO_2 may rise. Since carbon dioxide equilibrates across the blood-brain barrier more quickly than bicarbonate, the resulting increase in CSF carbon dioxide may cause a paradoxical rise in CSF hydrogen ion concentration. This may perpetuate the hyperventilation, even though the peripheral stimulus has been reduced by restoration of blood hydrogen ion concentration to normal.

There is general agreement that, in otherwise uncomplicated cases of non-respiratory acidosis that have a reversible cause (e.g. ketoacidosis and some forms of lactic acidosis), bicarbonate administration should not be considered unless the arterial hydrogen ion concentration is unusually severe (e.g. >100 nmol/L). In acute kidney injury, however, bicarbonate administration may have a rapid and beneficial effect on dangerous hyperkalaemia, while renal replacement treatment is being put in place. If intravenous bicarbonate is to be given, it should be given as a series of small quantities (e.g. 50 mmol) in isotonic solution (unless there is a danger of fluid overload, when a hypertonic solution may be used); after each infusion, the patient should be reassessed clinically and arterial $[H^+]$ and PCO_2 measured.

The use of bicarbonate in the treatment of renal tubular acidosis (RTA) is discussed on p. 79.

Specific causes of non-respiratory acidosis

Ketoacidosis. Diabetic ketoacidosis is discussed in detail in Chapter 16. The primary abnormality is increased lipolysis and ketogenesis (production of acetoacetic and 2-hydroxybutyric (β -hydroxybutyric) acids), coupled with decreased utilization of these acids. In addition, dehydration may decrease the glomerular filtration rate (GFR) and impair renal hydrogen ion excretion. Some patients also have a lactic acidosis (see below). It is noteworthy that resolution of the acidosis and clinical improvement of conscious level can lag behind the correction of the hyperglycaemia and ketosis in this condition. As normovolaemia is restored by the administration of 0.9% saline, renal excretion of sodium together with ketoacid anions results in loss of a source of bicarbonate and there is a tendency to hyperchloraemic acidosis, which may not be so rapidly reversed.

Ketoacidosis can also occur in association with alcohol ingestion (see Chapter 40), typically in alcoholics one or more days after a drinking bout, so that ethanol may not be detectable in the blood at the time that acidosis becomes apparent. Ketoacidosis develops as a result of a combination of factors, which are summarized in Table 5.2. The term ketoacidosis is a relative misnomer

TABLE 5.2 Factors involved in the pathogenesis of alcoholic ketoacidosis

Factor	Mechanism(s)
Fasting	Decreased insulin and increased counter-regulatory hormones, ^a leading to: increased lipolysis increased ketogenesis decreased ketone utilization
Hypoglycaemia	Effects of fasting exacerbated
Dehydration	Increased counter-regulatory hormones Impaired renal ketone excretion Impaired insulin secretion ^b
Ethanol	Metabolism inhibits lipolysis but lipolysis is stimulated as ethanol is cleared from the blood

^aCounter-regulatory hormones include glucagon, adrenaline (epinephrine) and cortisol.

^bVia stimulation of α -adrenergic receptors by adrenaline (epinephrine).

when applied to this condition; the oxidation of ethanol to acetaldehyde and acetate increases the [NADH]/[NAD⁺] ratio and hence the ratio [2-hydroxybutyrate]/[acetoacetate], so that dipstick tests for urinary ketones (which do not detect 2-hydroxybutyrate) may be negative or only weakly positive.

In both types of ketoacidosis, significant excretion of the organic acids can occur in urine; at the lowest attainable urinary hydrogen ion concentration, some 50% of 2-hydroxybutyrate, but only 10% of acetoacetate, is in the form of the undissociated acid. Although this is advantageous in that it represents a route of buffered hydrogen ion excretion, the fact that the bulk of acetoacetate and 50% of 2-hydroxybutyrate are excreted as organic cations has adverse consequences. It entails obligatory excretion of sodium and potassium as the accompanying anions, thus contributing to their deficits, and also removes substrate for the later regeneration of bicarbonate by oxidation of these anions. It is of interest that, in patients with ketoacidosis with poor renal function (in whom less excretion of organic anions will occur), the extent of the fall in plasma bicarbonate concentration at the time of admission tends to match the increase in the anion gap and in the plasma concentrations of organic acid anions; in contrast, in patients with good renal function, the anion gap is less and plasma chloride concentration is increased. A third, but rare, cause of ketoacidosis is severe starvation.

The non-enzymatic decarboxylation of acetoacetate to acetone consumes hydrogen ions, but adds to the burden of carbon dioxide to be excreted.

Other acid–base disturbances associated with alcohol. A range of other acid–base disturbances can occur in association with alcohol ingestion. In normal subjects, the severe hypoglycaemia that can occur when ethanol is taken after a period without food, can be accompanied by a mild lactic acidosis. In chronic alcoholics, severe

vomiting may cause a non-respiratory alkalosis, and continued ethanol ingestion can cause a lactic acidosis.

In patients with cirrhosis, alcohol can precipitate hepatic encephalopathy, which is often associated with respiratory alkalosis, as a result of stimulation of the respiratory centre by ammonia and other nitrogenous toxins.

Lactic acidosis. Lactate is formed from pyruvate as the end-product of glycolysis. Normal venous plasma lactate concentration is 0.6–1.2 mmol/L, tending to the lower end of this range during fasting and to the higher end after meals. Concentrations of up to 10 mmol/L can occur during severe physical exercise, but they fall rapidly once exercise ceases. Lactate is generated by glycolysis, principally in skeletal muscle, brain, erythrocytes, the skin and the gut, and is disposed of by gluconeogenesis in the liver and kidneys, thereby providing an important source of glucose, and by complete oxidation.

Lactic acid is a strong acid and is virtually completely dissociated at normal physiological hydrogen ion concentrations. Thus, the generation of lactate is always accompanied by equimolar generation of hydrogen ions. Buffering or other compensatory processes may prevent a significant rise in hydrogen ion concentration so that hyperlactataemia is not always associated with acidosis, but if the lactate concentration is greater than about 5 mmol/L, the hydrogen ion concentration is usually elevated.

Lactic acidosis can occur as a result of either excessive lactate formation or decreased lactate disposal or a combination of both. It is conventionally divided into type A (hypoxic) lactic acidosis, and type B, in which hypoxia is not the primary event. Causes of type B lactic acidosis include exposure to drugs, toxins and other chemicals, severe liver disease and certain inherited metabolic diseases. Some of the more important causes of lactic acidosis are indicated in **Box 5.2**. This list is not meant to be exhaustive; many rare conditions that have been reported to cause lactic acidosis and some conditions that only occasionally cause it, have not been included.

BOX 5.2 Some causes of lactic acidosis

Type A

- Shock (haemorrhagic, cardiogenic etc.)
- Severe hypoxia
- Severe exercise

Type B

- Associated with the ingestion of drugs etc.
 - Biguanides (especially phenformin)
 - Ethanol
 - Fructose administration (also sorbitol)
- Associated with inherited metabolic diseases
 - Fructose 1,6-diphosphatase deficiency
 - Pyruvate carboxylase deficiency
 - Pyruvate dehydrogenase deficiency
 - Glucose 6-phosphatase deficiency (glycogen storage disease type I)
- Other causes
 - Liver failure
 - Thiamin deficiency
 - Metabolic myopathies etc.

Type A lactic acidosis is the commoner variety. It is due primarily to an increase in lactate formation, as a consequence of tissue hypoxia, resulting, for example, from cardiogenic or haemorrhagic shock. Approximately half the lactate is produced by the gut. Decreased disposal of lactate is also important. Hepatic uptake and metabolism of lactate may be decreased in acidosis and when perfusion is decreased. The effects of acidosis on the cardiovascular system (negative inotropism and vasoconstriction) may further impair tissue perfusion (so that in the gut, a vicious spiral may develop, whereby splanchnic vasoconstriction induced by acidosis itself contributes to the acidosis). Although an acidosis-induced right shift in the oxyhaemoglobin dissociation curve may promote oxygen delivery to tissues, type A lactic acidosis can become self-perpetuating unless vigorous measures are undertaken to reverse the disturbance and treat the underlying cause. The prognosis is poor and appears to be related directly to the blood lactate concentration, mortality exceeding 80% with a concentration >9 mmol/L. The specific treatment of the acidosis, for example by the judicious administration of bicarbonate, has been described above. Measures to correct the underlying cause are more important.

The mechanisms of lactate accumulation in type B lactic acidosis vary. The classic descriptions of this condition relate to phenformin-induced lactic acidosis; phenformin is a biguanide formerly used for treating type 2 diabetes mellitus. It causes lactic acidosis primarily as a result of decreased utilization of lactate for gluconeogenesis, but increased production may also contribute. Lactic acidosis frequently leads to shock, and increased production may then become the predominant mechanism. Metformin, also a biguanide, has low risk of causing lactic acidosis except in patients with renal impairment, in whom it should not be used.

The inherited metabolic diseases shown in Box 5.2 are a useful paradigm to illustrate the mechanisms that can lead to the accumulation of lactate (Fig. 5.5). Two reactions in the glycolytic pathway are not simply reversible for gluconeogenesis. Deficiency of both fructose 1,6-diphosphatase and pyruvate carboxylase, the enzymes that catalyse these steps in gluconeogenesis, are associated with lactate accumulation, particularly when production is increased as, for example, during exercise. In glucose 6-phosphatase deficiency, glucose 6-phosphate released from glycogen cannot be converted to glucose and so is metabolized through the glycolytic pathway to lactate, particularly when glycogenolysis is stimulated, for example during fasting. Pyruvate dehydrogenase is responsible for converting pyruvate to acetyl coenzyme A, which subsequently enters the tricarboxylic acid cycle: its deficiency decreases pyruvate, and hence lactate, oxidation.

The association between alcohol and lactic acid, mentioned above, is not confined to ethanol. Methanol and ethylene glycol are metabolized by alcohol dehydrogenase and the reactions consume NAD^+ , which is then not available for the lactate dehydrogenase reaction. Particularly with ethylene glycol poisoning, other acids (e.g. glycolic and glyoxylic acids) also contribute to the acidosis (it should, however, be noted that glycolate may cause an apparent elevation of lactate concentration in lactate assays employing lactate oxidase).

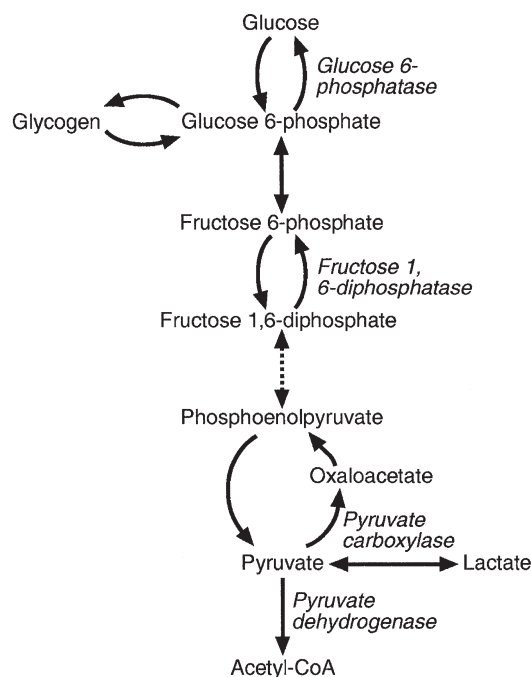


FIGURE 5.5 ■ Key reactions in glycolysis/gluconeogenesis. Inherited metabolic diseases affecting the enzymes indicated can cause lactic acidosis, by blocking gluconeogenesis (fructose 1,6-diphosphatase and pyruvate decarboxylase deficiencies), pyruvate oxidation (pyruvate dehydrogenase deficiency) or by increasing glycolysis (glucose 6-phosphatase deficiency). Intermediate reactions have been omitted for clarity.

Hyperlactataemia is common in patients with severe acute hepatic failure, but in many such patients the predominant acid–base disturbance is not acidosis, but alkalosis, either respiratory (as a result of direct stimulation of the respiratory centre by toxins) or non-respiratory, thought to be caused by a decrease in ureagenesis such that the consumption of hydrogen ions during the oxidation of the carbon skeletons of amino acids is not balanced by hydrogen ion formation.

The naturally occurring isomer of lactic acid in man is the L-isomer. Lactic acidosis due to accumulation of the D-isomer has been reported in patients with blind intestinal loops and short bowel syndrome, resulting from overproduction by the altered gut flora. The diagnosis may be missed, since D-lactate is not measured by the usual assays based on lactate dehydrogenase.

Dilutional (expansion) acidosis. Excessive infusion of isotonic saline solution can cause a mild acidosis. Expansion of the extracellular fluid (ECF) volume leads to a dilutional decrease in bicarbonate concentration, but, more importantly, volume expansion decreases renal bicarbonate reabsorption. This form of acidosis is self-correcting when ECF volume is allowed to return to normal.

Acidosis in renal disease. Acidosis is common in patients with renal disease. In acute kidney injury, acidosis is frequently multifactorial. Patients are often shocked and may have increased acid production, while the failure

of urine production prevents renal acid excretion. In chronic kidney disease, the urine can usually be maximally acidified – that is, the hydrogen ion concentration can be raised (the pH lowered) to the same extent as can occur in normal individuals (exceptions include renal disease particularly affecting the medulla, for example chronic pyelonephritis, in which acidification may be impaired). In the early stages, the development of hyperphosphataemia results in an increase in the amount of phosphate filtered in each nephron, and phosphate reabsorption is decreased secondarily to increased secretion of parathyroid hormone. This may allow the excretion of a normal load of hydrogen ion buffered by phosphate. However, although individual nephrons may secrete a normal or even increased amount of ammonium, overall ammonium excretion decreases early in chronic renal disease and, as a consequence, systemic acidosis develops. Nevertheless, severe acidosis is unusual in chronic kidney disease, with plasma bicarbonate concentration rarely falling below 10–12 mmol/L. This is a result of extensive buffering of hydrogen ions by bone, a factor of importance in the pathogenesis of renal osteodystrophy. Uraemic acidosis is an example of a type of acidosis in which alkali treatment is beneficial, though usually with oral

calcium carbonate rather than sodium bicarbonate in order to reduce the risk of sodium overload and also bind phosphate.

Impairment of renal tubular function is an important cause of acidosis. Three distinct syndromes of renal tubular acidosis (RTA) have been described: type 1 or distal RTA (sometimes referred to as ‘classic’ RTA); type 2 or proximal RTA, and type 4 or hyperkalaemic RTA. Both type 1 and type 4 owe their pathogenesis to defective mechanisms in the distal tubules: they are sometimes called hypokalaemic and hyperkalaemic distal RTA, respectively. Their principal characteristics are summarized in Table 5.3.

Normal urinary acidification requires the generation of nearly a 1000-fold gradient in hydrogen ion concentration between the blood and the lumen of the nephron, and type 1 RTA can be due to a failure to maintain this gradient because secreted hydrogen ions diffuse back into the tubular cells or due to defective hydrogen ion secretion. It can be inherited as an isolated defect (autosomal dominant) or be acquired. Unusually, for a systemic acidosis, there is associated potassium wasting, leading to hypokalaemia, partly because more potassium is excreted to maintain electrochemical neutrality and partly because the impairment of hydrogen ion excretion impairs

TABLE 5.3 Some causes and features of renal tubular acidosis

Type	Pathogenesis	Causes	Features
1 (distal)	Impaired ability to generate [H ⁺] gradient between blood and tubular lumen	Inherited idiopathic Autoimmune disease: Systemic lupus erythematosus: Primary biliary cirrhosis Sjögren syndrome Multiple myeloma Drugs: Amphotericin Lithium Non-steroidal anti-inflammatory drugs (NSAIDs) Renal disease: Analgesic nephropathy Sickle cell nephropathy Nephrocalcinosis	Minimum urine pH >5.5 Hypokalaemia Hypocitraturia Hypercalciuria Nephrocalcinosis Renal calculi Rickets Osteomalacia
2 (proximal)	Impaired reabsorption of filtered bicarbonate	Inherited idiopathic Metabolic disease: Tyrosinosis Cystinosis Hereditary fructose intolerance Wilson disease Drugs and toxins: Lead, mercury, acetazolamide Miscellaneous: Multiple myeloma Amyloidosis Sjögren syndrome	Minimum urine pH 4.5 if plasma [HCO ₃ ⁻] low Hypokalaemia Glycosuria Uricosuria Amino aciduria Hyperphosphaturia Hypophosphataemia Rickets Osteomalacia Polyuria Thirst
4 (hyperkalaemic)	Hypoaldosteronism	Adrenal failure Renal disease: Diabetic nephropathy Interstitial nephropathy Drugs: Indometacin Spironolactone	Minimum urine pH >5.5 ^a Hyperkalaemia Low plasma [renin] Low plasma [aldosterone]

^a Lower urinary pH may be achieved in ‘rate’ defects, see text for details.

sodium reabsorption and may lead to secondary aldosteronism. Buffering of acid in bone contributes to hypercalciuria and there is a risk of nephrocalcinosis. Type 1 RTA is typically treated with modest amounts of oral sodium bicarbonate (typically 2–3 mmol/kg body weight/day; cf. type 2 RTA, below); potassium supplements, citrate and thiazide diuretics (which increase proximal tubular sodium – and thus bicarbonate – reabsorption by inducing volume contraction) may also be helpful. The principal aim of treatment is to allow normal growth in children and to correct the bone disease and reduce the risk of renal damage in adults.

Type 2 RTA is less common than type 1. It is due to a failure of proximal tubular bicarbonate reabsorption, which leads to bicarbonaturia. The fall in plasma bicarbonate concentration causes a systemic acidosis, but as the amount of filtered bicarbonate falls, the amount presented to the proximal tubules for reabsorption falls, and bicarbonate may be completely reabsorbed. Since the distal acidification mechanism is normal, excretion of an acidic urine is possible, albeit only at the expense of a systemic acidosis. Plasma bicarbonate concentration is typically 15–20 mmol/L. Type 2 RTA is usually associated with other abnormalities of proximal tubular function, for example glycosuria, amino aciduria, hypercitraturia and phosphaturia; it can be a feature of inherited diseases (classically, the Fanconi syndrome caused by cystinosis) or be acquired. Plasma potassium concentration is usually normal or only slightly decreased. Hypercalciuria is not a feature of type 2 RTA. However, metabolic bone disease (rickets in children, osteopenia in adults) occurs frequently. It is, in part, a direct consequence of the acidosis; defective synthesis of calcitriol and hypophosphataemia may also be contributory factors. Correction of the acidosis in type 2 RTA requires the ingestion of much larger quantities (typically 5–15 mmol/kg body weight/day) of bicarbonate than are required in RTA type 1. Appropriate treatment is particularly important in children, to optimize growth (the condition is in any case rare in adults). Bicarbonate replacement leads to bicarbonaturia and potassium wasting; potassium replacement is also required.

Renal tubular acidosis type 4 can be subdivided into two types. So-called ‘rate defects’ (which occur more frequently) are associated with decreased secretion of aldosterone, resulting from either primary adrenal disease or renal disease in which renin secretion is impaired (e.g. diabetic nephropathy, tubulointerstitial nephropathy). They can also be seen in syndromes of aldosterone resistance, in patients being treated with spironolactone and in patients treated with non-steroidal anti-inflammatory drugs (in whom it is probably as a result of a decrease in renin secretion secondary to decreased renal prostaglandin synthesis). The acidosis is due, in part, to reduced hydrogen ion secretion secondary to decreased sodium reabsorption, and usually responds to treatment with small doses of a synthetic mineralocorticoid together with bicarbonate, diuretics or ion exchange resins to correct the hyperkalaemia. However, there is also decreased ammonium excretion, and the fact that this, and the acidosis, may be corrected if the plasma potassium concentration is lowered without giving aldosterone suggests that the hyperkalaemia itself contributes to the

acidosis. The other subtype comprises ‘voltage defects’, in which the secretion of hydrogen and potassium ions in the cortical collecting ducts is reduced as a result of a decrease in the usual negative intratubular electrical potential. This can occur with structural damage to the nephrons (e.g. sickle cell nephropathy), drugs (lithium, triamterene and amiloride, among others) and a decrease in the supply of sodium to the distal nephron as a result of avid proximal absorption (e.g. in hepatic cirrhosis). The urine pH can be reduced below 5.5 in rate defects, but not in voltage defects. In contrast to type 2 RTA, where acidosis is often severe, the acidosis in both subtypes of 4 RTA is often mild (plasma bicarbonate concentration typically 18–21 mmol/L), and the hyperkalaemia is often more remarkable.

If a diagnosis of renal tubular acidosis is not clear from the clinical findings and the simple laboratory measurements outlined above, tests of urinary acidification (for distal RTA) or bicarbonate reabsorption (for proximal RTA) may be required (see Chapter 9). Investigations that may be of value in the diagnosis of type 4 RTA include assessment of adrenal function and demonstration of a failure of plasma renin and aldosterone concentrations to respond to oral furosemide.

Respiratory acidosis

This is a consequence of carbon dioxide retention. In health, the rate and depth of respiration are adjusted precisely, so that the rate of excretion matches the rate of formation. Carbon dioxide excretion is a complex process, involving the transport of carbon dioxide in the blood to the pulmonary capillaries, diffusion into the alveoli and ventilation. Ventilation is controlled by the medullary respiratory centre, which receives input from peripheral and central chemoreceptors. Carbon dioxide retention can occur as a consequence of malfunction of either the excretory mechanism or its control. Some of the causes of respiratory acidosis are indicated in [Box 5.3](#).

Compensatory responses in respiratory acidosis

Buffering. The role of erythrocytes in converting carbon dioxide to bicarbonate and buffering the hydrogen ions produced has been discussed above. This process is extremely effective: in health, the arteriovenous difference in hydrogen ion concentration is only ~3 nmol/L. It is salutary to examine the efficacy of this process in more detail. In the absence of any buffering, were an increase in PCO_2 sufficient to cause an increase in bicarbonate concentration of only 1 mmol/L to occur (acutely, an increase of little more than 1 kPa is required to do this), it would be expected that hydrogen ion concentration would increase by the same amount, that is, 1 mmol/L. The fact that the increase is only a matter of a few nanomoles is almost entirely due to buffering by haemoglobin. Buffering of hydrogen ions by haemoglobin takes place rapidly, but the amount of haemoglobin available to buffer hydrogen ions is limited. The maximum increase in plasma bicarbonate concentration that occurs in acute respiratory acidosis is of the order of 4 mmol/L. In chronic carbon dioxide retention, buffering by other

BOX 5.3 Some causes of respiratory acidosis**Defects in the control of respiration**

- CNS depression
 - Anaesthetics
 - Narcotics
 - Severe hypoxia
- CNS disease
 - Trauma
 - Stroke
- Neurological disease
 - Spinal cord lesions
 - Poliomyelitis
 - Guillain–Barré syndrome
 - Motor neuron disease
 - Neurotoxins

Defects in respiratory function

- Mechanical
 - Myasthenic syndromes
 - Myopathies
 - Thoracic trauma and deformities
 - Pneumothorax, pleural effusion
- Pulmonary disease
 - Restrictive defects (extensive fibrosis, pulmonary oedema, infiltrative tumours)
 - Obstructive defects (chronic bronchitis, emphysema, severe asthma)
 - Impaired perfusion (massive pulmonary embolism)

intracellular buffers occurs, and, as discussed below, increased renal ammonium excretion plays an important (albeit indirect) part in controlling the hydrogen ion concentration. Buffering of hydrogen ions by bone occurs to a lesser extent in chronic respiratory acidosis than in metabolic acidosis.

Hyperventilation. Although the increase in PCO_2 will stimulate the respiratory centre, the underlying disease will mean that the respiratory apparatus is unable to respond adequately to this stimulus. Therapeutic measures to improve respiratory function may lower PCO_2 , but in chronic carbon dioxide retention, this may have undesirable consequences, as is discussed in the section on treatment, below.

Renal hydrogen ion excretion. In sustained carbon dioxide retention, renal bicarbonate reabsorption is maximal and phosphate is excreted almost entirely in the dihydrogen form. There is also a marked increase in urinary ammonium excretion. This has the effect of compensating for the increase in hydrogen ion formation from carbon dioxide, and may even restore blood hydrogen ion concentration to normal. It is accompanied by a further increase in plasma bicarbonate concentration (in addition to that generated directly by the erythrocyte mechanism). Although this is usually considered to be a consequence of increased renal hydrogen ion excretion (since bicarbonate is generated *pari passu* with hydrogen ions in renal tubular cells), it is probably also partly a result of diversion of ammonium from ureagenesis, requiring decreased buffering by bicarbonate of hydrogen ions produced during this process. Practically, it is important

to appreciate that this compensation evolves over several days of carbon dioxide retention. If an attempt is made to reduce PCO_2 rapidly, for example by artificial ventilation, temporary persistence of the compensatory process may result in the development of an alkalosis (Post-hypercapnic alkalosis, see p. 83).

There is a limit to which changes in renal acid excretion and in ammonium metabolism can compensate for an increase in PCO_2 ; if this rises above 8 kPa, arterial hydrogen ion concentration will always be increased.

Biochemical characteristics of respiratory acidosis

The cardinal features of respiratory acidosis are an increased blood PCO_2 and a high, or high-normal, hydrogen ion concentration; bicarbonate concentration is increased. The hydrogen ion and bicarbonate concentrations for any PCO_2 depend on the extent of the compensatory increase in renal hydrogen ion and ammonium excretion (see above). In an acute disturbance, the increase in bicarbonate is only of the order of 2–4 mmol/L, even with massive increases in PCO_2 , but in compensated disturbances the increase is much greater.

Systemic effects of respiratory acidosis

In patients with respiratory acidosis, the manifestations of the underlying disorder and of hypoxaemia, if present, usually dominate the clinical findings, but effects due to acidosis and to hypercapnia may also be present. Hypoxaemia causes breathlessness, cyanosis and drowsiness. The consequences of acidosis have been discussed above. The effects of hypercapnia are seen predominantly in the central nervous and cardiovascular systems.

The neurological effects of hypercapnia cover a spectrum from anxiety and confusion through to impaired consciousness and coma. Particularly in chronic carbon dioxide retention, headache, papilloedema, extensor plantar responses and myoclonus may occur. Most of these effects are due to the increased cerebral blood flow that is a consequence of the vasodilatory action of carbon dioxide.

Systemic vasodilatation also occurs, but the cardiac output is increased so that blood pressure is usually maintained. The skin is warm and arterial pulses are bounding. The acidosis may cause venous constriction and chronic hypoxaemia may cause pulmonary hypertension and cor pulmonale, rendering the patient very susceptible to pulmonary oedema should intravenous fluids be given injudiciously.

Management

The logical management of respiratory acidosis is to treat the underlying cause and thus restore PCO_2 to normal. This may not be possible and, in chronic carbon dioxide retention, if compensatory processes have restored the blood hydrogen ion concentration to normal or near normal, it may not be necessary. In practice, the management of respiratory disorders is usually dictated by the necessity to maintain an adequate arterial PO_2 . As mentioned above, rapid correction of an elevated PCO_2 in a patient with chronic carbon dioxide retention is

potentially dangerous. The compensatory changes can persist for several hours, or even a few days, and may cause the patient to become alkalotic.

This demonstrates the fact that the compensatory process in a respiratory acidosis can be regarded as the *physiological* generation of a non-respiratory alkalosis, although the patient's hydrogen ion concentration does not fall *below* normal, that is, the patient does not become frankly alkalotic, as long as the PCO_2 remains elevated.

Non-respiratory alkalosis

This disorder can develop because of excessive loss of hydrogen ions, decreased generation of hydrogen ions or exogenous alkali administration. Some of the causes are indicated in **Box 5.4** and some of the more important of these are discussed later.

Non-respiratory alkalosis is characterized by an increase in plasma bicarbonate concentration. Since this bicarbonate is filtered at the glomerulus and, therefore, available for urinary excretion, the persistence of a non-respiratory alkalosis implies that it is being perpetuated by inappropriate reabsorption of filtered bicarbonate. Indeed, in healthy subjects, it is very difficult to produce a sustained alkalosis by the administration of, for example, sodium bicarbonate, whether orally or intravenously, because the excess is excreted in the urine. In considering the body's responses to non-respiratory alkalosis, it is therefore necessary to take into account not only the

BOX 5.4 Some causes of non-respiratory alkalosis

Saline responsive (chloride depletion)

- Gastrointestinal causes
 - Vomiting
 - Gastric drainage
 - Congenital chloride-losing diarrhoea
- Exogenous alkali administration
 - Sodium bicarbonate, lactate, acetate especially if GFR reduced
- Contraction alkalosis
- Renal causes
 - Poorly reabsorbable anion therapy (e.g. carbenicillin)
 - Diuretics (secondary to ECF volume contraction)
 - Post-(chronic)hypercapnia, especially in congestive cardiac failure, nephrotic syndrome, cirrhosis, i.e. decreased 'effective intravascular volume'

Saline unresponsive

- Associated with hypertension
 - Primary aldosteronism
 - Cushing syndrome
 - Secondary aldosteronism
 - Liquorice, carbenoxolone
- Not usually associated with hypertension
 - Bartter, Gitelman syndromes
 - Refeeding after starvation
 - Potassium depletion^a (except in renal tubular acidosis, acetazolamide and pancreatic fistula)
 - Magnesium depletion^a

^aWhen diuretic induced, may respond to saline.

underlying cause, but also the factors responsible for its perpetuation. Three such factors appear to be important:

- extracellular volume contraction
- potassium deficiency
- mineralocorticoid excess.

These are discussed further below.

Compensation for non-respiratory alkalosis

Buffering. A fall in blood hydrogen ion concentration results in the release of buffered hydrogen ions and, in consequence, the blood bicarbonate concentration increases. The excess bicarbonate could be excreted by the kidneys but, as mentioned above, this process is impeded in sustained alkalosis.

Hypoventilation. In a systemic alkalosis, decreased stimulation of chemoreceptors would be expected to decrease the respiratory drive, leading to compensatory retention of carbon dioxide. However, any increase in PCO_2 will tend to stimulate respiration and lessen the extent of the acute compensatory response. With the passage of time, this response may increase, because the central respiratory centre appears to become less sensitive to an increase in PCO_2 . However, hypoventilation will tend to decrease arterial PO_2 and should the extent of this be so severe as to cause hypoxaemia, the stimulus of this will override any inhibitory effect of a low hydrogen ion concentration. Significant tissue hypoxia is, however, uncommon in the absence of respiratory disease: although alkalosis causes a left shift in the oxyhaemoglobin dissociation curve, chronically this is offset by a decrease in red cell 2,3-DPG, which reduces the affinity of haemoglobin for oxygen (see p. 90).

Renal bicarbonate excretion. As indicated above, persistence of a non-respiratory alkalosis implies continued and inappropriate renal bicarbonate reabsorption. This could be achieved by the combination of a fall in the glomerular filtration rate (GFR) with maintenance of normal rates of tubular bicarbonate reabsorption, or by enhanced tubular reabsorption with a normal GFR. In many patients with non-respiratory alkalosis, there is both a decrease in the GFR and increased bicarbonate reabsorption.

A decrease in ECF volume may lead to a decrease in the GFR. If this is associated with chloride deficiency, the requirement to maximize tubular sodium reabsorption may cause an obligatory increase in the reabsorption of filtered bicarbonate to maintain electrochemical neutrality, to the extent that the urine may become acidic. (Priority is given to the control of ECF volume over that of acid-base status.) In the majority of patients with non-respiratory alkalosis, correction of the alkalosis follows repletion of ECF volume by the infusion of an isotonic sodium chloride solution (hence 'saline-responsive' non-respiratory alkalosis).

Non-respiratory alkalosis is frequently associated with potassium deficiency, but the fact that, in many instances, the alkalosis can usually be corrected by volume expansion without replacement of potassium casts doubt on the precise role of potassium. However, potassium depletion can contribute to the maintenance of a non-respiratory alkalosis through an effect on bicarbonate reabsorption. Severe potassium depletion enhances

proximal bicarbonate reabsorption. Distal tubular sodium reabsorption takes place in exchange for potassium and hydrogen ions; particularly when there is enhanced distal sodium reabsorption, potassium depletion may result in an increased secretion of hydrogen ions into the tubular fluid.

The third factor that can maintain a non-respiratory alkalosis is an increase in mineralocorticoid activity. This promotes distal tubular sodium reabsorption and results in increased excretion of both potassium and hydrogen ions. The effect is potentiated by potassium depletion, increased distal sodium delivery (as with diuretic treatment) and by the presence of non-resorbable anions, which accentuate the negativity of the luminal aspect of tubular cells. Increased mineralocorticoid secretion can occur secondarily to ECF volume contraction or be primary, as in Conn and Cushing syndromes. In these two conditions, ECF volume is expanded and the alkalosis, unusually, is not corrected by saline infusion.

Biochemical characteristics of non-respiratory alkalosis

The blood hydrogen ion concentration is low and the bicarbonate concentration increased; respiratory compensation may increase PCO_2 , but not to more than about 8 kPa. Hypokalaemia is almost always present. Paradoxically, for the reasons explained above, the urine may be acidic and urinary potassium excretion increased.

Systemic effects of alkalosis

In general, the effects of alkalosis are the opposite of those of acidosis. The effects on the cardiovascular system are, however, less in alkalosis and infrequently of clinical consequence. Alkalosis is rarely sustained over a long period and there is no evidence of any adverse effects on bone.

Chronic non-respiratory alkalosis is frequently associated with potassium depletion and hypokalaemia. In addition to any contribution from the causative condition (e.g. gastric secretions contain approximately 10 mmol/L potassium), this is related to increased distal tubular secretion of potassium as a consequence of decreased hydrogen ion secretion. Thus, alkalosis can cause potassium depletion and potassium depletion sustain alkalosis. If sodium depletion is also present, the stimulation of maximal distal tubular sodium reabsorption will also contribute to potassium depletion.

Neuromuscular hyperexcitability is present frequently in patients with acute respiratory alkalosis, manifest as paraesthesia, muscle cramps and tetany. It is unusual in non-respiratory alkalosis, except when the hydrogen ion concentration falls rapidly, as has been reported in patients with chronic respiratory acidosis treated with mechanical ventilation. Grand mal convulsions have been reported in such patients, and alkalosis can precipitate a fit in patients with epilepsy. This is, at least in part, a result of buffering of hydrogen ion by plasma proteins, particularly albumin, which decreases in alkalosis, leading to increased binding of calcium to protein, thus lowering the plasma ionized calcium concentration.

Management of non-respiratory alkalosis

Management should be directed towards treatment of the underlying cause of the alkalosis, when possible. As mentioned above, treatment can also be directed towards the correction of any factors tending to sustain the alkalosis. Most patients will respond to expansion of the ECF volume with isotonic saline. The demonstration of a low urinary chloride concentration reliably predicts those patients who will respond to this treatment. This is frequently combined with potassium replacement, although, in many instances, the alkalosis can be corrected by the administration of saline alone, even if there is potassium depletion (which may require correction in its own right).

Administration of saline is inappropriate (and potentially dangerous) in patients with saline-unresponsive causes of non-respiratory alkalosis. Management must be directed towards the underlying cause, for example removing the source of excessive mineralocorticoid secretion (or blockade of mineralocorticoid action) or replacement of potassium or magnesium, as appropriate.

Specific causes of non-respiratory alkalosis

Loss of gastric acid. The most severe non-respiratory alkalosis is seen in patients losing unbuffered acid from the stomach because of either gastric drainage or prolonged vomiting, particularly in association with pyloric stenosis, which prevents the concomitant loss of alkaline secretions from the proximal small intestine. The acid is hydrochloric acid so that patients become chloride depleted. The hydrogen ions are derived from carbonic acid, *pari passu* with bicarbonate. Initially, renal excretion of the excess bicarbonate may prevent the development of alkalosis, but with ECF volume contraction, the requirement to maximize renal sodium reabsorption in the face of hypochloreaemia necessitates increased reabsorption of sodium with bicarbonate. Indeed, in severe cases, renal bicarbonate reabsorption may be complete in spite of the high plasma concentration, resulting in the excretion of an acidic urine. Potassium is lost in gastric fluid, but increased aldosterone secretion may result in significant loss of potassium in the urine as well, exacerbating the potassium depletion and the alkalosis. Thus all three of the factors mentioned above can be involved in the maintenance of the alkalosis. The alkalosis usually responds to re-expansion of the ECF volume with isotonic saline; isovolaemic infusion of saline prevents the development of alkalosis in patients whose gastric fluid is being drained (e.g. because of postoperative ileus).

Post-hypercapnic alkalosis. It has been pointed out that the rapid lowering of a chronically elevated PCO_2 can result in the compensatory processes causing a frank non-respiratory alkalosis. If PCO_2 falls to normal over a few days, bicarbonate is excreted and an alkalosis does not develop, but if this occurs in patients with contraction of the ECF volume, for example as a result of diuretic treatment, alkalosis may become apparent and persist. It can be treated by administration of isotonic saline, though with caution because these patients frequently have secondary cardiac disease.

Mineralocorticoid excess. Syndromes of actual or apparent mineralocorticoid excess are almost invariably associated with non-respiratory alkalosis, for reasons that have been outlined above. Extracellular fluid volume is increased and patients are usually hypertensive. Expansion of the ECF should oppose the effect of potassium depletion on proximal bicarbonate reabsorption, and it is likely that the alkalosis is maintained largely by increased distal reabsorption.

Miscellaneous. Several other conditions can cause non-respiratory alkalosis as a result of increased renal excretion of hydrogen ions, but are not associated with hypertension. They include Bartter and Gitelman syndromes, and magnesium and potassium depletion. Refeeding after starvation is also sometimes associated with transient non-respiratory alkalosis, the cause of which is unclear, but which may be perpetuated by concomitant hypovolaemia and avid renal sodium reabsorption.

Respiratory alkalosis

Respiratory alkalosis is a consequence of the rate of excretion of carbon dioxide exceeding the rate of production, leading to a decrease in PCO_2 . This is usually due to stimulation of the respiratory centre: the stimulus may be toxic, reflex, psychogenic or related to the presence of an intracranial lesion. The exception is mechanical ventilation, when normal respiratory control is overridden. It is a common abnormality in the critically ill. Some of the causes of respiratory alkalosis are indicated in [Box 5.5](#).

Compensatory responses in respiratory alkalosis

Buffering. In acute respiratory alkalosis, the fall in PCO_2 causes a decrease in hydrogen ion concentration and a slight fall in bicarbonate. Other buffers release hydrogen ions, so tending to counter the effect of the fall in PCO_2 ; some of these hydrogen ions will combine with bicarbonate, causing its concentration in the blood to fall further. A new steady state can be attained rapidly and can persist for approximately six hours, after which the effect of changes in renal hydrogen ion metabolism become detectable.

BOX 5.5

Some causes of respiratory alkalosis

- Voluntary hyperventilation
- Mechanical ventilation
- Reflex hyperventilation
 - Decreased pulmonary compliance
 - Disease affecting chest wall
 - Irritative lesions of the air passages
- Other stimuli to respiratory centre
 - Cortical influences (pain, fever, anxiety etc.)
 - Local disease (trauma, tumours)
 - Drugs and toxins (salicylate poisoning, hepatic failure)
 - Hypoxaemia (ascent to altitude, right-to-left shunts, pulmonary disease, carbon monoxide poisoning etc.)
- Non-respiratory acidosis (during recovery)

Hypoventilation. Correction of a respiratory alkalosis is only possible if the rate of excretion of carbon dioxide can be restored to normal. The fact that an alkalosis develops (other than with mechanical ventilation) indicates that the inhibitory effect of the decrease in PCO_2 on respiration is being overwhelmed by whatever stimulus is causing the hyperventilation.

Renal hydrogen ion excretion. If a low PCO_2 persists for more than a few hours, decreased renal generation of bicarbonate (for which carbon dioxide is a substrate) will decrease urinary acidification and effect further compensation for the alkalosis, further lowering the plasma bicarbonate concentration.

Biochemical features of respiratory alkalosis

The cardinal feature of an acute respiratory alkalosis is a decrease in arterial PCO_2 , a decrease in hydrogen ion concentration and a small decrease in bicarbonate concentration, though not to less than about 18 mmol/L. In a chronic respiratory alkalosis, renal compensation may result in arterial hydrogen ion concentration being only marginally decreased, while the bicarbonate concentration falls further, but not to less than about 12 mmol/L. The finding of bicarbonate concentrations less than these values suggests the additional presence of a non-respiratory acidosis.

If the stimulus to hyperventilation is hypoxaemia, arterial hydrogen ion concentration may be affected predominantly by a resulting non-respiratory acidosis. The interpretation of measured acid-base parameters in mixed disorders of hydrogen ion homeostasis is discussed in a later section.

Systemic effects of respiratory alkalosis

The manifestations of the underlying disorder often predominate in patients with respiratory alkalosis. In acute hypocapnia, cerebral vasoconstriction reduces cerebral blood flow, and light-headedness, confusion, impaired intellectual function, syncope and seizures may occur. In patients with sickle cell disease, hypocapnia has been recorded as causing strokes, presumably as a result of cerebral hypoxaemia engendered by the vasoconstriction. Perioral and peripheral paraesthesiae are common, in part at least, because of a fall in ionized calcium concentration. These features usually remit if the hypocapnia persists.

In contrast, cardiovascular changes can occur with both acute and chronic hypocapnia. They include an increase in heart rate, non-specific chest pain or even frank angina.

Mild hypokalaemia can occur with respiratory alkalosis; plasma phosphate concentration is often significantly reduced.

Management

Where possible, treatment should be directed at the underlying cause. In a psychogenically induced acute respiratory alkalosis, rapid symptomatic relief may be obtained by getting the patient to rebreathe from a paper bag.

If the alkalosis is severe, and the nervous or cardiovascular features are giving cause for concern, it may occasionally be necessary to sedate the patient or prevent the hyperventilation by resorting to mechanical ventilation, always being sure that adequate oxygenation is maintained.

The interpretation of acid–base data

Many approaches to the diagnosis of disorders of hydrogen ion homeostasis have been promulgated. These may involve the mathematical manipulation of data, or the plotting of data on a diagram that shows the ranges of the variables in the various disorders. Three points are worth emphasizing in this context. First, although the use of diagrams may facilitate the rapid diagnosis of an acid–base disturbance, it cannot add to the information that is available from an analysis of the data on which the diagrams are based. Second, diagnosis should be based on measured variables and not on secondarily derived data. Third, acid–base variables must always be assessed in their clinical context: this is particularly important in mixed disorders, in which results may be indistinguishable from those that can occur during the physiological compensation of single disorders.

The most logical acid–base diagram is thus a graph of hydrogen ion concentration against partial pressure of carbon dioxide, as shown in Figure 5.6. This shows the zones in which combinations of PCO_2 and $[H^+]$ occur in the various disorders that have been described. If a pair of data for a patient falls outside one of these areas, it suggests that there is a mixed acid–base disorder. Such diagrams may be useful not only to help in defining the acid–base disorder in an individual patient, but also, when serial plots are made, to follow the response to treatment.

It will be apparent that diagnosis of the type of acid–base disorder requires only knowledge of PCO_2 and $[H^+]$;

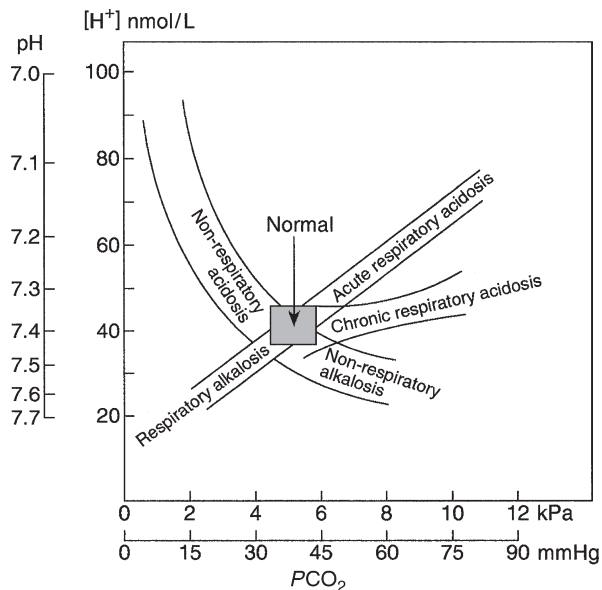


FIGURE 5.6 ■ The relationship between arterial hydrogen ion concentration ($[H^+]$) and partial pressure of carbon dioxide (PCO_2). The ranges of values typical of the major disturbances of hydrogen ion homeostasis are shown. Note that, in acute respiratory disturbances, the variation of $[H^+]$ with PCO_2 is rectilinear.

this must follow from the fact that, as shown by Equation 7, bicarbonate concentration depends on PCO_2 and $[H^+]$ and cannot vary independently of them.

In Figure 5.6, the plot of $[H^+]$ against PCO_2 is rectilinear in acute respiratory disorders. This fact makes it possible to diagnose acid–base disorders without recourse to diagrams. The slope of the plot is such that, in an acute respiratory disturbance (that is, before there has been time for significant compensation to occur), for every 1 kPa change in PCO_2 , $[H^+]$ would be expected to change by 5.5 nmol/L, an increase in PCO_2 increasing $[H^+]$ and a decrease in PCO_2 decreasing $[H^+]$ by this amount.

Thus, in a patient whose PCO_2 is abnormal, it is possible to calculate whether the observed $[H^+]$ is what it would be expected to be if the respiratory disturbance were acute, that is, in the absence of any compensation. Suppose, for example, a patient has a PCO_2 of 7.5 kPa and $[H^+]$ of 52 nmol/L – plotting these figures on Figure 5.6 suggests that the patient has an acute respiratory acidosis. But this can also be appreciated by calculation: assuming a normal PCO_2 of 5.3 kPa and $[H^+]$ of 40 nmol/L, an acute increase in PCO_2 to 7.5 kPa would be expected to increase $[H^+]$ to $[40 + (5.5 \times 2.2)] = 52$ nmol/L, as is observed. Were the observed $[H^+]$ to be higher, it would indicate that the acidosis had a non-respiratory as well as a respiratory component. Were the observed $[H^+]$ to be lower than expected from the PCO_2 , it would indicate the presence of a non-respiratory alkalosis in addition to the respiratory acidosis.

In this latter instance, however, an alternative explanation could be that the respiratory acidosis had been partly compensated. As has been indicated, the compensatory processes in disturbances of hydrogen ion homeostasis can be regarded as the generation of opposing disturbances. Indeed, in respiratory disturbances, in which the development of compensation lags behind the development of the primary disorder, the presence and nature of the compensation is made explicit if the primary abnormality is corrected rapidly, leaving the persisting compensatory process as the sole acid–base abnormality (see Post-hypercapnic alkalosis, above).

As previously mentioned, it is important to appreciate that examination of PCO_2 and $[H^+]$, or of data derived from them, in isolation from clinical information and other, independent, laboratory data, may not permit differentiation between a compensated disorder of hydrogen ion homeostasis or the simultaneous existence of two primary disorders having opposing effects on hydrogen ion concentration. In the descriptions of the major acid–base disorders, the maximum extent to which compensation can occur was indicated, and if data lie outside these limits then a mixed disturbance must be present; however, such a disturbance can be present even if the data lie within the limits of normal compensation.

The approach to the interpretation of acid–base data described above is essentially physiological; it does not require the use of derived data which, as explained in an earlier section, do not provide additional, independent, information.

An alternative approach, which is essentially physicochemical in nature, is provided by the strong ion theory, originally developed by Stewart (1978). This was based

on in vitro studies of the dissociation of water (and hence hydrogen ion concentration) in the presence of various solutes and carbon dioxide, and led to the conclusion that $[H^+]$ is determined by three independent variables. These are the 'strong ion difference' (SID, the sum of the concentrations of sodium, potassium, magnesium and calcium ions less than that of the sum of the concentrations of chloride, sulphate and organic acid anions); the total concentration of weak acids, i.e. phosphate and proteins (A_T) and the partial pressure of carbon dioxide.

Strong ion difference is further differentiated into 'apparent' SID (SID_a) and 'effective' SID (SID_e). The former is provided by $([Na^+] + [K^+]) - [Cl^-]$; the latter is the sum of $[HCO_3^-]$ and non-bicarbonate buffers, that is, the anionic equivalences of albumin and phosphate. The difference between the two is the 'strong ion gap' (SIG). The normal value of SIG is zero. The various quantities are either measured or read from nomograms based on the solutions of a series of complex equations.

This approach defines additional primary acid–base disorders in addition to the four described above. There are two respiratory disorders (as with the physiological approach), but in non-respiratory acidosis, although SID_e is always low, this can occur with a high SIG (as in lactic acidosis) or a normal SIG (as with bicarbonate loss); in non-respiratory alkalosis, SID_e is high and SIG normal, but both acidosis and alkalosis can also be due to abnormal value of A_T (hyperalbuminaemic acidosis and hypoalbuminaemic alkalosis, respectively). The contribution of differences in plasma protein concentration is not considered to be of significance in the physiological approach (and indeed, changes in plasma albumin concentration in vivo do not affect $[H^+]$).

Although the strong ion theory has its proponents, it has not been widely adopted. It is considerably more complex than the physiological approach; it is derived entirely from observations made in vitro; it does not provide any quantitative data concerning the compensatory responses to primary disorders and, although there are some data suggesting that it better predicts outcome in critically ill patients than traditional approaches, most studies have failed to do this. (The interested reader will find further information and discussion of this topic in the Further reading section, below.)

Mixed disorders of hydrogen ion homeostasis

Mixed acid–base disorders, that is, disorders with both respiratory and non-respiratory components, occur frequently. Some examples are shown in Box 5.6. They can be classified according to whether the component disturbances are additive or opposing in their effects on hydrogen ion concentration.

The presence of a mixed disorder can be inferred from any change in hydrogen ion concentration not being as predicted from the PCO_2 . In a mixed respiratory and non-respiratory acidosis, PCO_2 will be elevated, but $[H^+]$ will be higher than would be predicted from the PCO_2 . In a mixed respiratory and non-respiratory alkalosis, PCO_2 will be decreased, but $[H^+]$ will be lower than predicted. In each case, the plot of $[H^+]$ and PCO_2

BOX 5.6 Some causes of mixed acid–base disorders

Additive disorders

- Non-respiratory acidosis + respiratory acidosis
 - Respiratory failure
 - Cardiac arrest
 - Poisoning, e.g. with ethanol, methanol
- Non-respiratory alkalosis + respiratory alkalosis
 - Vomiting and congestive cardiac failure
 - Vomiting or diuretic therapy and, for example, hepatic failure or pneumonia

Counterbalancing disorders

- Non-respiratory acidosis + respiratory alkalosis
 - Salicylate poisoning
 - Septicaemia ± renal failure
 - Acute hepatic failure, hepatorenal syndrome
 - Ketoacidosis and pneumonia
- Non-respiratory alkalosis + respiratory acidosis
 - Diuretic therapy or vomiting and chronic obstructive airways disease
 - Severe potassium depletion
- Non-respiratory acidosis + non-respiratory alkalosis
 - Vomiting and renal failure
 - Diuretic therapy and ketoacidosis
 - Severe vomiting in ketoacidosis

Triple disorders

- See text

on the acid–base diagram (Fig. 5.6) will fall in the regions between those shown for the appropriate non-respiratory and acute respiratory disorders.

The diagnosis of mixed disorders in which the component disturbances have opposite effects on hydrogen ion concentration is less straightforward. As discussed above, the problem is that the changes in $[H^+]$ and PCO_2 may be exactly the same as may be found as a result of physiological compensation. This stems from the fact that the compensatory process itself disturbs acid–base homeostasis, albeit advantageously, tending as it does to restore $[H^+]$ towards normal. Usually, careful consideration of the clinical findings and other data that may be available will allow the correct diagnosis to be made. It should also be appreciated that the efficacy of compensatory processes is limited, particularly in non-respiratory disorders. Compensation restores the hydrogen ion concentration *towards* normal, but restoration *to* normal is only seen with mild chronic respiratory acidosis and chronic respiratory alkalosis. With more severe disturbances, compensation is incomplete. Overcompensation does not occur so that, for example, a patient with a slightly elevated $[H^+]$ and a low PCO_2 , must have a partially compensated non-respiratory acidosis and cannot have compensated respiratory alkalosis.

Since the causes of respiratory acidosis and alkalosis are retention and excessive excretion of carbon dioxide, respectively, these two disorders cannot coexist. However, several mechanisms can be responsible for the development of non-respiratory disorders, and the

existence of a process producing an acidosis need not exclude the presence of one producing an alkalosis. Thus, a patient could have renal failure (a cause of acidosis) and be vomiting excessively (a cause of alkalosis). Whether the patient is actually acidotic or alkalotic will depend on which process predominates, but examination of $[H^+]$ and PCO_2 alone will not reveal this type of mixed disturbance. It may only be inferred from clinical observations coupled with other biochemical data – particularly, for example, the anion gap – or become apparent when the predominant disorder is treated appropriately and apparent ‘over-swing’ occurs, for example the patient, having had a mild non-respiratory acidosis, becomes severely alkalotic.

Exceptionally, triple acid–base disorders can occur. Patients have been described with a severe non-respiratory alkalosis (usually due to prolonged vomiting), accompanied by a respiratory acidosis or alkalosis, in whom a high anion gap suggests a component of non-respiratory acidosis (e.g. a lactic acidosis). The $[H^+]$ (low) and PCO_2 (also low) in such patients may suggest a partially compensated acute respiratory disturbance, but the clinical features will be inconsistent with this as the sole diagnosis.

In practice, because mixed acid–base disorders frequently develop in patients who are severely ill, the management of the acid–base disorder per se will usually be less important than that of the underlying illness. Indeed, the management of any acid–base disorder should include measures to correct the cause where possible, though, as has been indicated above, specific measures relating to the acidosis or alkalosis may also be necessary.

TISSUE OXYGENATION

Introduction

The process whereby atmospheric oxygen is made available to mitochondria, where the oxidation of carbon and hydrogen releases energy, is a complex one, depending on adequate alveolar ventilation and function, pulmonary and tissue blood flows, and the ability of the blood itself to take up oxygen in the alveoli and release it to tissues. Tissue oxygenation can be compromised by disease affecting any of these functions. Until relatively recently, the only readily available index of tissue oxygen supply was the arterial partial pressure of oxygen, PaO_2 . This measurement is still regarded as essential, but it has a number of limitations. It requires access to arterial blood, either by direct puncture or through an indwelling catheter; blood samples must be collected with considerable care and analysis performed without delay, but, perhaps most importantly, measurement of PaO_2 provides incomplete information on oxygen transport. It is a measure of partial pressure, not of the oxygen *content* of blood nor the *delivery* of oxygen to tissues. However, although tissue oxygenation depends upon factors other than PaO_2 , maintenance of an adequate PaO_2 is a prerequisite for normal tissue oxygenation. This is illustrated in the following sections, which discuss the transport of oxygen from the inspired gas to the tissues, where it is used for oxidative metabolism.

Pulmonary function

The lungs have two principal functions: to transfer oxygen from inspired gas to the blood, and to remove carbon dioxide from the blood to the expired gas. As discussed above, the latter is critical to hydrogen ion homeostasis, and is compromised by hypoventilation. Hypoventilation invariably leads to decreased oxygen uptake into the blood, but, as will be seen, impaired oxygenation can occur in the absence of hypoventilation.

Alveolar ventilation

The partial pressure of oxygen in arterial blood, PaO_2 , depends on the alveolar oxygen tension (PAO_2), which is in turn dependent on the fraction of the inspired gas comprising oxygen (FiO_2), the arterial carbon dioxide tension ($PaCO_2$), the respiratory quotient (RQ), atmospheric barometric pressure (P_B) and the partial pressure of water vapour (PH_2O) such that:

$$PaO_2 = FiO_2 \times (P_B - PH_2O) - PaCO_2 / RQ \quad (29)$$

This is the alveolar gas equation. The respiratory quotient depends on the relative proportions of free fatty acids, carbohydrate and protein being used as energy substrates by the tissues, but it varies through only a small range, even in disease. Alveolar air is always saturated with water so that PH_2O is constant. It follows that PAO_2 can increase significantly as a result of either an increase in P_B (requiring a hyperbaric chamber) or FiO_2 (requiring the administration of oxygen) or a decrease in $PaCO_2$ (requiring an increase in ventilation). Only the last two are readily available to treat patients with a low PAO_2 . Inspection of Equation 29 will indicate that the effect of a change in $PaCO_2$ on PAO_2 will be greater if FiO_2 is normal than if it is raised; an increase in $PaCO_2$ (usually accepted as being equal to $PaCO_2$) can significantly decrease PAO_2 at normal values of FiO_2 , but it will have a lesser effect if the FiO_2 is increased.

Oxygen uptake into blood

The continuous delivery of mixed venous blood of low oxygen content to the alveolar capillaries, and diffusion of oxygen down the concentration gradient from the alveolar space to the blood, result in a constant tendency for PAO_2 to fall that is prevented by the delivery of oxygen to the alveoli by ventilation. In healthy young individuals, oxygen diffuses readily from alveoli into the plasma and PaO_2 is usually only about 1 kPa less than PAO_2 , that is (breathing room air at sea level) about 13.3 kPa. The gradient increases with age and the approximate normal value is given by $(PAO_2 - PaO_2) = (0.06 \times \text{age in years})$ kPa so that it may reach nearly 4 kPa in healthy 60-year-olds. Pulmonary disease that impairs diffusion is a potential cause of an increase in $(PAO_2 - PaO_2)$ and hence a decrease in PaO_2 , but the nature of the alveoli is such that any impairment of diffusion must be considerable before it affects PaO_2 at rest.

The maintenance of a normal PaO_2 also requires a normal relationship between the perfusion of alveoli and

their ventilation: the effects of a disturbance in this relationship (ventilation–perfusion imbalance) are further considered below.

In health, a small proportion of the blood reaching the lungs from the tissues bypasses the alveoli and does not take part in gas exchange (a right-to-left shunt). This is because the bronchial veins drain directly into the pulmonary veins, while some blood that perfuses the myocardium drains directly into the cavity of the left ventricle. Any increase in shunting due to a pathological process will tend to decrease PaO_2 .

The final factor that can affect PaO_2 is the oxygen tension in the blood reaching the lungs, that is, mixed ('mixed' because this blood is derived from all the veins draining into the right side of the heart) venous PO_2 or PvO_2 . If this is low, increased alveolar oxygen transport will be necessary to allow maximal PaO_2 . In health, increased tissue oxygen requirements (as, e.g. during exercise) result in a fall in PvO_2 , and the increased oxygen requirement is met by hyperventilation. Tissue oxygen requirements are often increased in disease, for example as a result of sepsis or the metabolic response to trauma, but, at the same time, the physiological responses leading to hyperventilation may be attenuated. Other factors that can contribute to a low PvO_2 include decreased oxygen saturation, anaemia and a decrease in cardiac output, all of which, as will be discussed below, can decrease the delivery of oxygen to tissues.

In summary, a low PaO_2 , that is, hypoxaemia, can be caused by any of the following:

- a decrease in PAO_2 (whether due to a decrease in the proportion of oxygen in the inspired gas, an increase in $PACO_2$ or a decreased barometric pressure)
- hypoventilation
- decreased diffusion
- imbalance of ventilation and perfusion
- an increase in shunting
- a decrease in PvO_2 .

Hypoxaemia caused by hypoventilation or a decrease in PAO_2 or PvO_2 can be distinguished from the other causes by calculating the alveolar–arterial oxygen gradient ($PAO_2 - PaO_2$). PAO_2 can be calculated from arterial blood gas measurements using the alveolar gas equation. The gradient is increased when hypoxaemia is a result of impaired diffusion, shunting or imbalance between ventilation and perfusion. In these conditions, $PaCO_2$ is often normal. However, when, as in hypoxaemia caused by alveolar hypoventilation, $PaCO_2$ is increased, ($PAO_2 - PaO_2$) is typically normal.

Hypoxaemia becomes recognizable clinically when it is sufficiently severe to cause central cyanosis. This requires a deoxygenated haemoglobin concentration of more than 50 g/L (saturation <75% at normal haemoglobin concentrations). In patients with severe anaemia, the low haemoglobin may result in significant hypoxaemia being present without there being discernible cyanosis; conversely, patients with polycythaemia may have central cyanosis, despite a relatively normal oxygen saturation.

Peripheral cyanosis will always be present if there is central cyanosis, but peripheral cyanosis can occur in the absence of central cyanosis as a consequence of reduced peripheral circulation.

The role of haemoglobin in oxygen transport

Although the amount of oxygen present in physical solution in the blood is directly related to PaO_2 (0.225 mL/L per kPa), only a small amount of oxygen is carried in this way (at normal PaO_2 , approximately 3 mL per litre of plasma). Oxygen is principally transported in the blood bound to haemoglobin. One gram of haemoglobin can bind 1.34 mL of oxygen when fully saturated. The normal PaO_2 of arterial blood is approximately 13.3 kPa, at which haemoglobin is approximately 97% saturated. Total arterial oxygen content (CaO_2) is given by the sum of the dissolved and haemoglobin-bound fractions:

$$CaO_2 = ([Hb] \times SaO_2 / 100) \times 1.34 + (0.23 \times PaO_2) \quad (30)$$

where [Hb] is the haemoglobin concentration (g/L) and SaO_2 is the percentage saturation of haemoglobin with oxygen. At sea level, CaO_2 is normally about 200 mL/L. SaO_2 can be measured using a co-oximeter; these instruments also measure haemoglobin concentration and thus allow calculation of arterial oxygen content, which provides far more information than PaO_2 alone. Modern blood gas analysers often incorporate a co-oximeter, and such instruments are widely used in intensive therapy units. Arterial oxygen content can also be measured non-invasively (transcutaneously) using a pulse oximeter. It is beyond the scope of this book to discuss in detail the analytical principles utilized in these instruments, but a summary is provided in a later section (see p. 91).

The familiar sigmoid relationship between Hb saturation and oxygen tension (Fig. 5.7) has a number of important clinical consequences. First, a considerable reduction in PaO_2 below normal has little effect on the amount of oxygen carried in the blood. Saturation only falls below 90% when PaO_2 falls below 8 kPa. If PaO_2 falls further, however, oxygen saturation (and thus the amount of oxygen carried) decreases sharply. A further consequence is that, because haemoglobin is saturable, increasing PaO_2

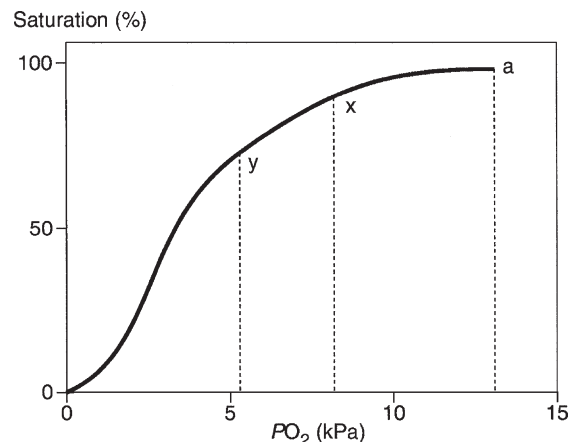


FIGURE 5.7 ■ The oxygen–haemoglobin dissociation curve. At a normal arterial PO_2 (point 'a'), haemoglobin is about 97% saturated ($SaO_2 = 97\%$). Point 'y' represents the normal venous PO_2 , at which haemoglobin is 75% saturated ($SvO_2 = 75\%$). A decrease in arterial PO_2 to 8 kPa (point 'x') is associated with a decrease in SaO_2 to only 90%, but a further decrease will have a proportionately greater effect on SaO_2 and thus on arterial oxygen content.

above that necessary to provide complete saturation has relatively little effect on arterial oxygen *content*, since only the small fraction present in solution is significantly increased.

The effects of pulmonary disease on oxygen uptake into blood

The effects of alveolar hypoventilation and impaired diffusion on pulmonary function have been described above. Two other functional defects, shunting and ventilation–perfusion imbalance, can also have profound effects on the oxygenation of blood.

Shunting

Haemoglobin is only about 97% saturated in normal arterial blood. This is, in part, because of physiological shunting. In some conditions (e.g. lobar pneumonia, pulmonary oedema, adult respiratory distress syndrome), some alveoli become filled with fluid and do not take part in gas exchange, although they are still perfused with blood. Shunting is thereby increased and this leads to arterial hypoxaemia. Atelectasis (collapse of a lung or part of a lung so that it is not aerated) has the same effect. Under such circumstances, increasing PAO_2 by increasing FiO_2 has little effect on overall PaO_2 . This is because no gas exchange will be taking place in the alveoli that are acting as shunts, while haemoglobin in blood leaving normally functioning alveoli will already be fully saturated with oxygen. Increasing FiO_2 will increase only the small proportion of oxygen held in physical solution in this blood, and this will have little effect on total arterial oxygen content.

Ventilation–perfusion imbalance

The fact that haemoglobin is not fully saturated in normal arterial blood is also related to an imbalance between alveolar ventilation and perfusion. At rest, ventilation is about 4.2 L/min and pulmonary blood flow is about 5.5 L/min, so that the overall ventilation–perfusion ratio (\dot{V}/\dot{Q} ratio) is approximately 0.8. However, this ratio is not uniform throughout the lungs, ranging between the approximate limits 0.5 and 3.0. Some alveoli are better ventilated than they are perfused ($\dot{V}/\dot{Q} > 1$), so that a proportion of ventilation is ‘wasted’ and the effect is an increase in ‘dead space’; in others, \dot{V} and \dot{Q} are in balance. In both cases, effective oxygenation of the blood takes place. However, in those alveoli in which perfusion exceeds ventilation ($\dot{V}/\dot{Q} < 1$), complete oxygenation of the blood is impossible. Because haemoglobin is saturable, oxygen transport in the well-ventilated and normally ventilated alveoli cannot compensate for decreased oxygenation of blood in poorly ventilated alveoli. Many pulmonary diseases, notably chronic obstructive pulmonary disease and interstitial pulmonary disease, give rise to an increase over the normal imbalance of ventilation and perfusion. However, in contrast to disease in which significant shunting occurs, increasing the FiO_2 in such conditions, because it will increase oxygen transport in poorly ventilated alveoli, can increase overall PaO_2 .

Differential effects of pulmonary disease on $PaCO_2$ and PaO_2

It will be instructive at this point to compare the effects of respiratory disease on $PaCO_2$ and PaO_2 . Carbon dioxide is transported in blood effectively in physical solution (albeit mostly in the form of bicarbonate). In contrast to the oxyhaemoglobin dissociation curve, the curve relating the carbon dioxide content of blood to partial pressure is almost linear over the physiological range. As a result, a significant change in $PaCO_2$ will always cause a significant change in the carbon dioxide content of blood. In contrast, as has been seen, considerable changes in PaO_2 can occur with little effect on the blood’s oxygen content.

$PaCO_2$ is determined by alveolar ventilation: in health, any increase in carbon dioxide formation (e.g. during exercise) can be matched by an increase in excretion and PaO_2 is not affected. In contrast, in pulmonary disease, if $PaCO_2$ is increased, then PaO_2 will always be decreased. Thus, if hypoventilation causes hypercapnia, it will also cause hypoxaemia. However, hypercapnia is not always present in patients who are hypoxaemic. Carbon dioxide can diffuse between the blood and the alveolar air more readily than oxygen; defects in diffusion rarely cause hypercapnia, although they can cause hypoxaemia. Shunting causes a decrease in PaO_2 , but this will stimulate ventilation, increasing the excretion of carbon dioxide from ventilated alveoli and preventing a rise in $PaCO_2$ or even causing hypocapnia. Only with extensive shunting will the $PaCO_2$ be increased.

In pulmonary disease causing an increase in \dot{V}/\dot{Q} imbalance, there is a tendency for PaO_2 to fall, for the reasons discussed above; this will stimulate respiration but, while the increased ventilation of well-perfused alveoli cannot compensate for the impaired oxygenation, it can increase carbon dioxide excretion and $PaCO_2$ fall. With more severe disease, however, hypoxaemia will be accompanied by hypercapnia.

Respiratory failure, defined by PaO_2 being < 8 kPa or $PaCO_2$ being > 7 kPa, is thus divided into two types: type I, in which PaO_2 is low and $PaCO_2$ is normal or low (typically seen in condition such as pneumonia, pulmonary oedema and acute lung injury, where lung tissue is damaged) and type II, in which PaO_2 is low but $PaCO_2$ is high (typically seen in chronic obstructive pulmonary disease and conditions causing hypoventilation).

Oxygen transport to tissues

Oxygen delivery

The transport of oxygen to tissues depends not only on adequate transfer of oxygen from the inspired gas to the alveolar capillaries, but also on an adequate cardiac output. The total amount of oxygen delivered by the cardiopulmonary apparatus (DO_2) is given by the product of oxygen content (CaO_2) and cardiac output (CO):

$$DO_2 = CaO_2 \times CO \quad (31)$$

However, the amount of oxygen available to tissues will depend upon local perfusion and the affinity of haemoglobin for oxygen, which determines how readily oxygen can be released.

Oxygen uptake

Oxygen is taken up into tissues because their PO_2 is lower than that of the blood, and oxygen diffuses down its concentration gradient. However, examination of the oxyhaemoglobin dissociation curve indicates that the amount of oxygen taken up will depend on the haemoglobin saturation (SO_2) at a given PO_2 . Mixed venous oxygen saturation (SvO_2) at rest is approximately 75%, corresponding to a PvO_2 of 5.3 kPa and permitting the uptake of 46 mL of oxygen from each litre of blood. Resting cardiac output is approximately 5 L/min, and total tissue uptake of oxygen is about 250 mL/min at rest. However, if the oxyhaemoglobin curve were to shift to the left, less oxygen would be released from haemoglobin for the same fall in pO_2 , whereas a right shift would increase the availability of oxygen.

Such shifts occur physiologically. An increase in PCO_2 or $[H^+]$ (both of which occur as blood traverses the capillary beds) causes a right shift, as does an increase in temperature. The position of the curve is also determined by the concentration of 2,3-DPG in erythrocytes. An increase in 2,3-DPG concentration (normally about 4 mmol/L of red cells) causes a right shift. This occurs in chronic hypoxia and thus facilitates oxygen uptake by tissues. These effects are illustrated in Figure 5.8. The position of the oxyhaemoglobin saturation curve can be defined by P_{50} , the partial pressure of oxygen at which haemoglobin is 50% saturated. It is normally about 3.7 kPa.

Hypoxia

Tissue hypoxia can be due to a disturbance occurring at any stage in the delivery of oxygen to the cells where it is utilized (Table 5.4), to increased demand or to deficient

TABLE 5.4 Factors affecting tissue oxygenation and causes of hypoxia

Factor	Cause of hypoxia	Type of hypoxia
Inspired oxygen content (FiO_2)	Low inspired PO_2 Low barometric pressure (P_B)	Hypoxic
Alveolar oxygen tension (PAO_2)	Hypoventilation Increased oxygen consumption Increased $PACO_2$	Hypoxic
Arterial oxygen tension (PaO_2)	Venous-arterial shunting V/Q imbalance Impaired diffusion capacity	Hypoxic
Oxygen-carrying capacity of blood	Low haemoglobin Low P_{50}	Anaemic
Blood supply to tissues	Low cardiac output Vasoconstriction	Stagnant
Impaired utilization of oxygen	Metabolic poisons	Histotoxic

oxygen uptake. The rational treatment of hypoxia obviously depends on knowing the cause.

Measurement of oxygen delivery to tissues

This requires determination of arterial oxygen content, preferably from measurements of haemoglobin and SaO_2 , and of cardiac output. SaO_2 can be calculated from PaO_2 , using an assumed value for P_{50} , but direct measurement, by oximetry, is more reliable. It may be possible to infer the adequacy of cardiac output from clinical observation, but it can be measured directly by the thermodilution

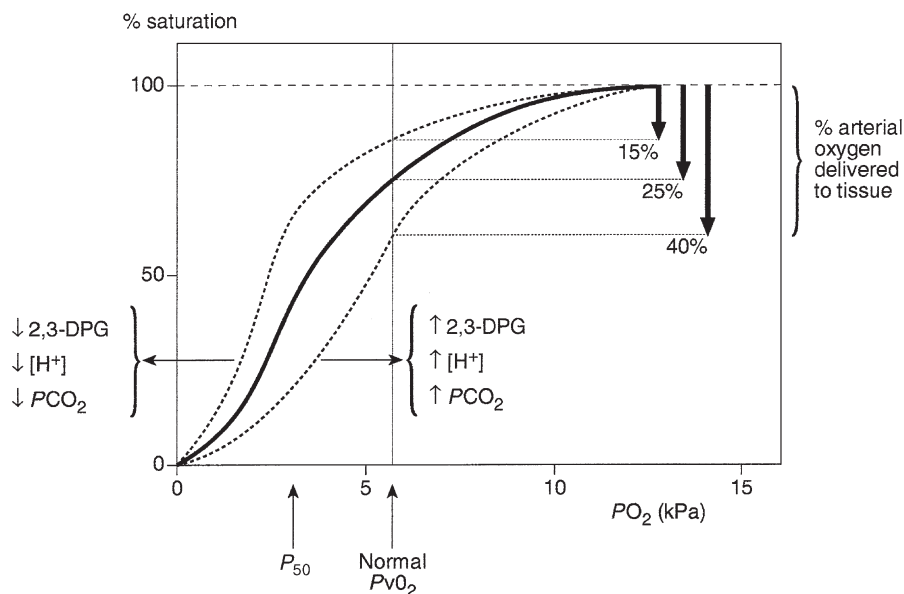


FIGURE 5.8 ■ The oxygen–haemoglobin dissociation curve. An increase in PCO_2 , $[H^+]$ or red cell 2,3-diphosphoglycerate (2,3-DPG) causes a right shift in the curve, decreasing the affinity of haemoglobin for oxygen (P_{50} increased) and increasing the amount of oxygen available to the tissues. A decrease in PCO_2 , $[H^+]$ or red cell 2,3-DPG causes a left shift, which has the opposite effect.

method using a pulmonary artery (Swan–Ganz) catheter incorporating a thermistor.

Oximeters – instruments that measure the saturation of haemoglobin with oxygen – are of two types. Co-oximeters measure the optical absorbance of blood in vitro at multiple wavelengths; because oxyhaemoglobin and deoxygenated haemoglobin (and, indeed, carboxyhaemoglobin and methaemoglobin) have different absorbance characteristics, these instruments can be calibrated to quantitate these species separately, and thus to measure SaO_2 . Pulse oximeters are non-invasive. They have sensors that are usually placed on a thin part of the patient's anatomy (e.g. a finger tip or ear lobe) that direct light of two wavelengths (660 and 940nm) through the tissue to a detector that measures the relative absorbance and hence haemoglobin and oxyhaemoglobin. The influence of absorption by tissues and venous and capillary blood is minimized by measuring the *changing* absorption (hence pulse oximetry) that is due to haemoglobin in arterial blood alone. Reflectance pulse oximeters are also available, but are less widely used.

Pulse oximeters do not detect carboxyhaemoglobin and other haemoglobin variants and may thus give misleading results if these are present. For example, while the saturation may be normal (e.g. 97%), if 12% of the haemoglobin were to be in the form of carboxyhaemoglobin, the true percentage of oxyhaemoglobin would only be 85% of the total. Pulse co-oximeters that combine the two techniques and allow more reliable non-invasive measurement of SaO_2 have been developed.

Most blood gas analysers provide an estimate of SaO_2 based on measurements of $[H^+]$, PaO_2 and empirical equations. This approach relies on a number of assumptions that may not be valid in individual patients, with the result that these estimates can vary significantly from measured values, particularly if abnormal haemoglobins (e.g. carboxy-, methaemoglobin) are present. Direct measurements are to be preferred. Blood gas analysers may also provide a value for arterial oxygen content, but as with saturation, these are only estimates, not true measurements.

Detection of tissue hypoxia

Although the accumulation of lactic acid is a consequence of severe tissue hypoxia, the development of hyperlactataemia is a relatively late phenomenon, and by the time it is detectable, hypoxic tissue damage may already have occurred. Furthermore, it can occur for other reasons, for example impaired hepatic function, strenuous muscle contraction (not only in exercise but also due to rigors or convulsions) and as a result of improved perfusion of *previously* poorly perfused tissue ('washout' phenomenon).

The measurement of mixed venous oxygen saturation (SvO_2) is now widely used in the assessment of the critically ill. SvO_2 can be measured either in vitro in a blood sample taken through a pulmonary artery catheter or in vivo using a catheter containing fibre-optic bundles, which transmit light of appropriate wavelength to the blood and transmit reflected light back to

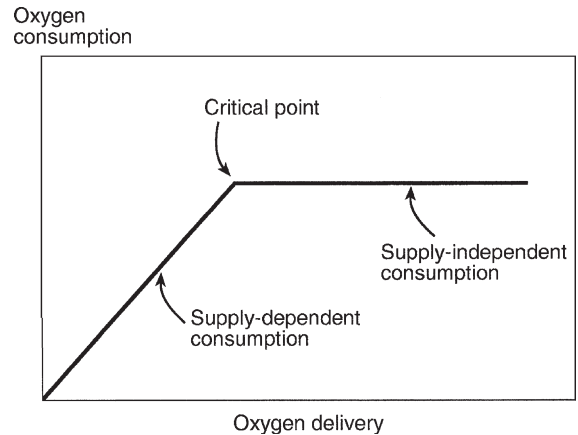


FIGURE 5.9 ■ Oxygen delivery to, and consumption by, tissues. Normally, consumption is independent of delivery; when delivery falls below a critical level, it results in decreased consumption.

a measuring device. It reflects the difference between arterial oxygen supply and tissue consumption. SvO_2 is normally of the order of 75%, but frequently falls to 50% or less when there is anaerobic metabolism, although such a fall can also be due to appropriately increased tissue uptake of oxygen or to a decrease in cardiac output. Because the measured SvO_2 is effectively a mean value of venous blood from all tissues, it can be affected by changes in the relative distribution of blood between different tissues. Measurements of SvO_2 can be misleading; in severe sepsis, reduced tissue oxygenation is sometimes associated with a high SvO_2 , as a result of impaired oxygen extraction and local arterio-venous shunting.

Oxygen delivery to tissues (approximately 16 mL/kg body weight/min at rest) normally exceeds demand, so that if delivery falls demand may still be met and consumption be independent of supply. However, if supply falls below a critical level (approximately 8 mL/kg per min), oxygen consumption becomes supply dependent (Fig. 5.9). If this occurs, it may be possible to improve oxygenation by the use of inotropes or red cell transfusion (to improve supply) or sedation (to reduce consumption).

Management of respiratory failure

It is beyond the scope of this book to discuss this topic in detail. Appropriate measures, depending on the cause, may include treatment of any airways obstruction, infection or pulmonary oedema, increasing FiO_2 , improving alveolar hypoventilation to decrease $PACO_2$, using inotropic drugs to stimulate cardiac output, increasing the oxygen-carrying capacity of the blood by red cell transfusion and the use of measures to increase tissue perfusion or reverse the effects of toxins. Hypophosphataemia is a sometimes unrecognized cause of respiratory muscle weakness, especially in the critically ill, as is malnutrition. In patients with chronic hypercapnia who are hypoxaemic, it has long been thought that increasing the FiO_2 is potentially dangerous. The rationale for this notion

is that acquired insensitivity of the respiratory centre to carbon dioxide can cause hypoxaemia to be providing the major stimulus to respiratory effort, and that its abolition would lead to respiratory arrest. However, hypoxaemia is a greater threat to life than hypercapnia, and this danger is now thought too have been exaggerated.

There are many techniques for improving alveolar hypoventilation, including the provision of invasive respiratory support through the use of a mechanical ventilators or even extracorporeal oxygenation techniques. Ventilators may be used in a variety of modalities, for example to take over completely a patient's breathing (e.g. when there is paralysis of the respiratory muscles or in deep coma leading to absent or poor respiratory effort) or to provide assistance to inspiratory efforts initiated by the patient. However, mechanical ventilation, although a potentially life-saving technique in critical hypoxia, is not without its disadvantages, including impaired cardiac filling (and hence a decrease in cardiac output); too rapid a fall in P_{aCO_2} ; depressing the normal respiratory drive; dilutional hyponatraemia due to inappropriate anti-diuresis, and a risk of mechanical damage to the airways and lungs (which may already be damaged). Oxygen in high concentrations is potentially toxic, and the lowest concentration compatible with reversing and preventing hypoxia should be used.

Ventilatory support is not synonymous with invasive mechanical ventilation. Patients able to breathe spontaneously may be helped by the maintenance of continuous positive airways pressure (CPAP) through the use of tight-fitting nasal or facial masks. The increased pressure aids gas entry during inspiration and reduces airways collapse during expiration.

CONCLUSION

Disorders of hydrogen ion metabolism and of tissue oxygenation are frequently encountered in clinical practice and may coexist, often because of a common aetiology,

but sometimes because of a direct effect of one upon the other. For both groups of disorders, a sound understanding of the physiological principles governing homeostasis in normal individuals is essential for correct diagnosis and effective management. These principles are essentially simple, and pertain no matter how complex any particular disturbance may appear.

Further reading

- Adrogue HJ, Madias NE. Renal tubular acidosis. In: Davison AM, Cameron JS, Grünfeld J-P et al. editors. Oxford textbook of clinical nephrology. 3rd ed. Oxford: Oxford University Press; 2005. p. 976–94. *A thorough review of these disorders.*
- Adrogue HJ, Gennan FJ, Gala JH et al. Assessing acid–base disorders. *Kidney Int* 2009;76:1239–47. *This review provides a succinct description and comparison of the approaches to assessing acid–base disorders (including the strong ion approach) with numerous references, convincingly arguing that the physiological approach is to be preferred.*
- Cohen RD, Woods HF. Disturbances of acid–base homeostasis. In: Warrell DA, Cox TM, Firth JD, editors. Oxford textbook of medicine. Oxford: Oxford University Press; 2010. p. 1738–51. *In addition to providing a detailed account of non-respiratory acid–base disorders, this chapter also emphasizes the importance of the liver in acid–base homeostasis.*
- Gentile M, Davies JD. Bedside monitoring of pulmonary function. In: Vincent J-L, Abraham E, Moore FA et al. editors. Textbook of critical care. 6th ed. Philadelphia: Elsevier Saunders; 2011. p. 279–87. *A clear explanation of the principles, use and disadvantages of pulse oximetry and other bedside techniques.*
- Marini J, Dries DJ. Principles of gas exchange. In: Vincent J-L, Abraham E, Moore FA et al. editors. Textbook of critical care. 6th ed. Philadelphia: Elsevier Saunders; 2011. p. 288–95. *A clearly written account of respiratory gas transport.*
- Palmer BP, Narins RG, Yee J. Clinical acid–base disorders. In: Davison AM, Cameron JS, Grünfeld J-P et al. editors. Oxford textbook of clinical nephrology. 3rd ed Oxford: Oxford University Press; 2005. p. 321–46. *A comprehensive account of non-respiratory disorders, with particularly good coverage of pathophysiology and extensive references.*
- Stewart PA. Independent and dependent variables of acid–base control. *Respir Physiol* 1978;33:9–26. *The first description of the strong ion theory.*
- Critical Care Medicine Tutorials. <http://www.ccmtutorials.com/rs/oxygen/page11.htm>; [Accessed 19.01.12]. *A succinct explanation of the concept of delivery dependent oxygen consumption.*

Calcium, phosphate and magnesium

Timothy Cundy • Andrew Grey • Ian R. Reid

CHAPTER OUTLINE

CALCIUM METABOLISM 93

- Biological role of calcium 93
- Distribution of calcium 93
- Calcium fluxes 94
- Regulation of calcium metabolism 95
- Biochemical assessment of calcium metabolism 99
- Hypercalcaemia 100
- Hypocalcaemia 105

PHOSPHORUS METABOLISM 109

- Distribution of body phosphorus 109
- Hyperphosphataemia 113
- Hypophosphataemia 115

MAGNESIUM METABOLISM 117

- Plasma magnesium 117
- Magnesium homoeostasis 117
- Hypomagnesaemia 118
- Hypermagnesaemia 121

CONCLUSION 121

APPENDICES 121

CALCIUM METABOLISM

Biological role of calcium

Calcium is a divalent cation with multiple roles in vertebrate physiology, which can be grouped as either structural or metabolic. Its structural role is in the skeleton where calcium deficiency leads to skeletal disease, either in the form of osteoporosis or osteomalacia. Its metabolic roles are more numerous.

The extracellular concentration of calcium ions influences the threshold for nerve action potentials, a high calcium raising the threshold and a low calcium having the opposite effect. The extracellular calcium concentration appears to alter the gating of sodium and potassium channels, possibly the result of calcium ions being attracted to and 'screening' the negative charge on the cell surface in the region of these channels.

Calcium has a key role in intracellular signalling. The depolarization of nerve or muscle cells, or the binding of a hormone or cytokine to its receptor on the surface of other types of cell, results in an increase in cytosolic calcium concentration via either one or both of two mechanisms: the influx of calcium through plasma membrane channels and the release of calcium from intracellular stores (e.g. the sarcoplasmic or endoplasmic reticulum). The latter pathway depends on receptor-activated hydrolysis of inositol phospholipids in the plasma membrane, leading to the release of inositol trisphosphate into the cytosol. This compound binds to a receptor on the endoplasmic reticulum resulting in the release of calcium into the

cytosol. The regulation of a variety of cell functions can then follow, some enzymes being directly affected by the cytosolic calcium concentration (e.g. protein kinase C) and others indirectly by the calcium receptor protein, calmodulin. These pathways are integral to muscle contraction, neuroendocrine secretion, cell metabolism and growth.

Distribution of calcium

The total body calcium in a human adult is approximately 1 kg, 99% of which is contained in the skeleton. Approximately 1% of skeletal calcium is freely exchangeable with calcium in the extracellular fluid (ECF). Calcium ions diffuse freely throughout the extracellular space, where their concentration is approximately 1.2 mmol/L. The plasma *ionized* calcium concentration is the same, but the plasma *total* calcium is approximately two-fold higher because of protein binding of calcium. Calcium also forms complexes with phosphate, citrate and bicarbonate in plasma and interstitial fluid (Table 6.1). It is the ionized calcium concentration that is physiologically important and closely regulated.

Calcium is principally an extracellular ion; cytosolic concentrations are of the order of 10^{-4} to 10^{-3} mmol/L. The low intracellular calcium concentration is necessary for calcium to function as an intracellular messenger. It is maintained by calcium pumps and exchangers on cell membranes. The endoplasmic reticulum and the mitochondria also have the capacity to remove calcium from the cytosol.

TABLE 6.1 Distribution of plasma calcium

Fraction	Total (%)
Ultrafilterable	53
Ionized	47
Complexed	6
Protein bound	47
Albumin	37
Globulin	10
Total (2.2–2.6 mmol/L)	100

Calcium fluxes

Three principal organs are involved in the body's handling of calcium: the gastrointestinal tract, bone and the kidneys (Fig. 6.1).

Gastrointestinal tract

Intestinal absorption of calcium is mediated by two mechanisms. One is an active transcellular process, regulated by calcitriol (1,25-dihydroxyvitamin D; $1,25(\text{OH})_2\text{D}$) in the duodenum. This involves uptake of calcium into the enterocyte by TRPV6 (membrane calcium channel transient receptor potential cation channel, subfamily V, member 6), followed by intracellular binding of calcium to the calcium-binding protein CaBP-9k, then energy-dependent transport of the calcium across the basolateral membrane via PMCA1b, an alternatively spliced transcript of plasma membrane Ca^{2+} -ATPase. 1,25-Dihydroxyvitamin D increases gene expression of TRPV6 and CaBP-9k, thus increasing

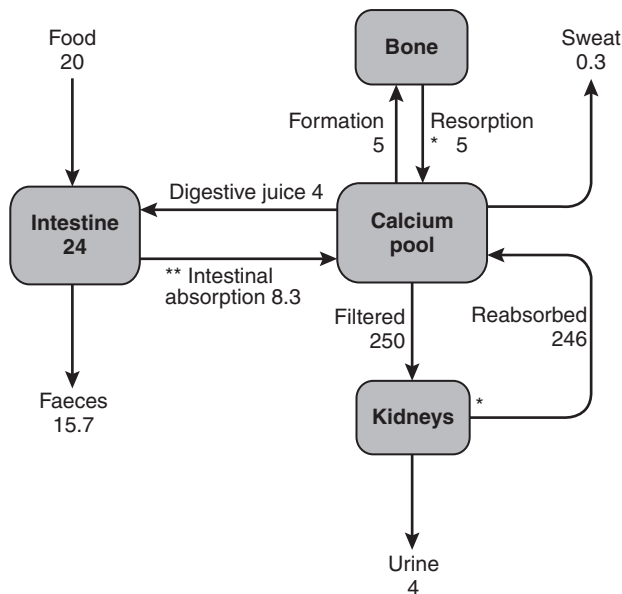


FIGURE 6.1 ■ Representative calcium fluxes (mmol/24h) in a healthy adult (70kg body weight) in zero calcium balance. The rapid exchange of calcium on bone surfaces and the exchange in soft tissues are not illustrated in this diagram. The major sites of action of parathyroid hormone (*) and 1,25(OH)₂D (**) are indicated. (Modified from Wilkinson R. Absorption of calcium, phosphorus and magnesium. In: Nordin B E C (ed). Calcium, phosphorus and magnesium metabolism. Edinburgh: Churchill Livingstone; 1976, with permission).

calcium absorption. Calcium is also absorbed passively throughout the small intestine and possibly in the colon. Because the duodenum is relatively short compared to the rest of the small bowel, it may account for less than half of the calcium absorbed at normal dietary intakes, despite its greater absorptive capacity per unit length. At low calcium intakes, active absorption predominates, but this is saturable, and at higher intakes, the change in net absorbed calcium in relation to any change in intake is relatively small and mediated by passive paracellular absorption.

The gut not only absorbs calcium but also secretes it in the digestive juices. The unabsorbed portion of this secreted calcium (approximately 2–3 mmol/24h) is known as the endogenous faecal calcium. Thus at very low intakes, the absorbed dietary calcium will be less than that lost from secretion in digestive juice and the net calcium absorption will be negative. Net calcium absorption is positive in healthy adults when their daily calcium intake is greater than 5 mmol.

The absorption of calcium is influenced by other dietary constituents. The presence of anions such as phosphate, oxalate (found in some fruit and vegetables) and phytate (found in some unrefined cereals) diminishes calcium solubility and thus net absorption. Gastrointestinal absorption of calcium tends to decline with age, but is increased during pregnancy and lactation. The principal regulator of intestinal absorption is 1,25(OH)₂D, and so either deficiency or excess of this hormone is associated with parallel changes in calcium absorption.

The average calcium intake of healthy adults consuming a Western diet is about 20 mmol/day, of which 20–40% is absorbed. The principal source of calcium in the Western diet is dairy products. There are significant amounts of calcium in some green vegetables such as spinach, though the bioavailability is lower.

Kidneys

The ionized and complexed fractions of plasma calcium are filtered at the glomeruli, amounting to approximately 250 mmol/24h. Of this, ~98% is reabsorbed, mostly in the proximal tubules. Calcium reabsorption in the proximal tubules is through a paracellular, therefore passive, process, which is closely linked with that of sodium and does not appear to be hormonally regulated. The fine-tuning of calcium excretion occurs in the distal parts of the nephrons where ~15% of the filtered load is reabsorbed. This section consists of the distal convoluted tubules, the connecting tubules and the initial portion of the cortical collecting ducts. In these distal sites, calcium reabsorption is active and occurs against an electrochemical gradient.

Active calcium reabsorption is a multistep process with initial passage of Ca^{2+} across the luminal membrane through the epithelial channel TRPV5 (membrane calcium channel transient receptor potential cation channel, subfamily V, member 5), then cytosolic diffusion bound to vitamin D-sensitive calcium-binding proteins (calbindins), and finally active extrusion across the opposite basolateral membrane by a Na^+ , Ca^{2+} -exchanger (NCX1)

and/or Ca^{2+} -ATPase (PMCA1b). This active calcium reabsorption is subject to hormonal regulation – principally by parathyroid hormone (PTH) and extracellular calcium itself, but possibly also by $1,25(\text{OH})_2\text{D}$, calcitonin, oestrogens and androgens.

Urine calcium excretion is increased in subjects consuming a high protein diet, because of acid production in the metabolism of sulphur-containing amino acids and sequestration of calcium by sulphate in the urine, resulting in inhibition of calcium reabsorption. Sodium intake can also influence calcium excretion by affecting proximal tubular sodium and calcium reabsorption (increased sodium intake increasing calcium excretion). Other factors that increase renal tubular calcium reabsorption are hypovolaemia, alkalosis and thiazide diuretics, whereas volume expansion, acidosis and loop diuretics (e.g. furosemide) have the opposite effect.

Bone

In the mature adult, the movement of calcium into bone equals its rate of efflux and bone mass remains constant. The situation is different in both growth and senescence. In spite of this constancy of adult bone mass, there is an active exchange of calcium between bone and the ECF. This can take place as a result of bone remodelling (see Chapter 31), or it can be accomplished by a process of mineral exchange between bone and the ECF, without local changes in bone matrix. Bone remodelling accounts for the changes in bone density that take place with ageing or disease. In the healthy adult, about 5% of the entire skeleton is remodelled in one year. In contrast, radioisotope studies have indicated that 1–2% of total body calcium can be exchanged between the bone and ECF over a period of several days. The precise mechanisms of this exchange and the factors influencing it are not known, but the quantities of calcium involved suggest that it may be an important part of normal calcium homeostasis.

Calcium losses in sweat are in the order of 0.3 mmol/24h. Breast milk has a high calcium content (~7.5 mmol/L) and a breast-feeding woman can lose 4–8 mmol/24h in her milk.

Regulation of calcium metabolism

Plasma calcium concentration is principally controlled by PTH and $1,25(\text{OH})_2\text{D}$. Calcitonin may also be regarded as a calcitropic hormone, although it has no clearly defined physiological function in humans. Many other hormones, growth factors and cytokines can influence bone metabolism.

Parathyroid hormone

Parathyroid hormone is an 84-amino acid, single-chain polypeptide secreted by the parathyroid glands. The gene is on chromosome 11 and consists of three exons encoding a peptide of 115 amino acids (pre-pro-PTH), which is cleaved to produce pro-PTH (90 amino acids) and subsequently the mature peptide (84 amino acids), before secretion. Most of the biological activity of PTH appears to reside in the 34 amino acids at the N-terminal end of the peptide.

The principal regulator of PTH secretion is the ECF ionized calcium concentration: low concentrations stimulate secretion and high concentrations inhibit it. This regulation is mediated by the calcium sensing receptor (CaSR), a cell membrane-bound, G-protein-coupled receptor. The relationship between PTH secretion and ionized calcium concentrations is not linear. Marked hysteresis is evident from experiments where ionized calcium concentrations are raised or lowered. The PTH concentration at any given concentration of plasma ionized calcium is lower when the ionized calcium concentration is rising than it is when ionized calcium is falling (Fig. 6.2). Plasma PTH concentrations exhibit a diurnal variation. They are stable during the afternoon and evening, but then rise by about 50% to peak at around 02.00 h and subsequently fall to values approximately 50% below afternoon values around 09.00 h. The zenith to nadir difference (in healthy men) is ~2.3 pmol/L, using an intact PTH assay.

Many other factors have been shown to influence PTH secretion, including $1,25(\text{OH})_2\text{D}$, adrenergic agonists, prostaglandins and magnesium, but whether any of these are physiologically important is debatable. Severe magnesium deficiency can produce a state of reversible hypoparathyroidism.

Parathyroid hormone binds to cell surface receptors in its target tissues. The PTH receptor has seven transmembrane domains, an extensive extracellular domain involved in hormone binding and an intracellular domain regulating intracellular signalling. In its two principal target tissues, bone and kidney, this results in activation of both adenylate cyclase (with production of cyclic adenosine

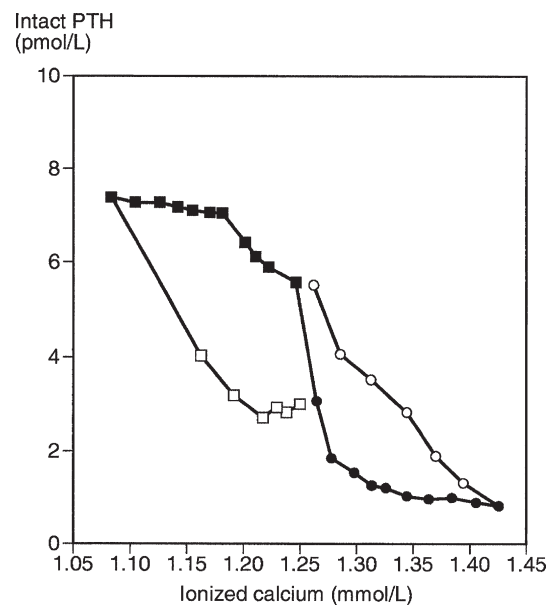


FIGURE 6.2 ■ Mean concentrations of plasma intact PTH in relation to plasma ionized calcium concentration in response to infusions of citrate (■) and calcium gluconate (●), and during recovery from hypocalcaemia (□) and hypercalcaemia (○). For any given ionized calcium concentration, the PTH concentration is lower when the ionized calcium is rising than when it is falling. (Data from Conlin P R et al. *Journal of Clinical Endocrinology and Metabolism* 1989; 69:593–599, with permission of the Endocrine Society).

monophosphate, cAMP) and phospholipase C (with production of inositol trisphosphate) and subsequent mobilization of intracellular calcium.

In bone, PTH receptors are present on cells of osteoblastic lineage. The osteoblasts in turn regulate osteoclast maturation through production of RANKL (receptor activator of nuclear factor κ -B ligand), which binds to the receptor, RANK, on pre-osteoclasts, leading to their maturation to osteoclasts. This process is modulated by osteoblast production of osteoprotegerin (OPG), which can also bind RANK-ligand, thus reducing its binding to RANK and so the stimulation of osteoclastogenesis.

Parathyroid hormone has three principal actions on the kidneys. It reduces proximal tubular reabsorption of phosphate, increases distal tubular reabsorption of calcium and increases the activity of the 25-hydroxyvitamin D 1α -hydroxylase enzyme in proximal tubular cells. It also decreases proximal reabsorption of bicarbonate, leading to a mild hyperchloraemic acidosis in states of PTH excess.

Measurement of circulating parathyroid hormone.

The development of assays for PTH has greatly simplified the investigation of patients with abnormal plasma calcium concentrations, but these assays do have limitations. The parathyroid gland secretes both the intact 84-amino acid peptide and inactive PTH fragments. Hepatic and renal metabolism of intact PTH results in the appearance of mid-region and carboxy-terminal peptides in the circulation. In vivo, the principal biologically active peptide is intact parathyroid hormone, which is present at a low concentration (1–5 pmol/L) and has a short half-life (2–4 min). However, mid-molecule and carboxy-terminal fragments have significantly longer half-lives and accumulate in concentrations 5–20 times greater than that of the intact peptide. These inactive fragments are cleared by the kidneys and their concentrations increase substantially in the presence of renal insufficiency. Although these fragments are not biologically active, they may be immunoreactive, if they cross-react with the particular antibodies used in the immunoassay. Assays that detect inactive fragments, therefore, report higher PTH concentrations than those that detect the intact molecule.

In recent years, two-site and immunoassays of 'intact' PTH, which purportedly measure only the biologically active intact peptide, have been widely used. However, there is evidence that these first-generation assays for intact PTH also detect non-(1–84)-PTH peptides, in particular PTH 7–84, which may have different biological actions from the full-length peptide. Assays based on antibodies directed at the first four amino acids of PTH (second-generation intact PTH assays) reveal that non-(1–84)-PTH peptides account for about 15% of circulating immunoreactive PTH (as detected by first-generation assays) in healthy patients, and at least 30% in patients with either primary or secondary hyperparathyroidism. There is also evidence that some non-intact peptides might be biologically active, possibly antagonizing the actions of the intact hormone. Whether assays based on antibodies directed at the extreme N-terminus of PTH offer important clinical advantage in the management of either primary hyperparathyroidism or renal bone disease

over the widely available first-generation intact PTH immunoassays remains uncertain.

Both mid-molecule and intact PTH assays show an increase in reference values with age, though this is more marked for the mid-molecule assays because of increased fragment accumulation secondary to the age-related decline in renal function. Because of the effect of renal function on measurements using mid-molecule and carboxy-terminal PTH assays, their interpretation must always take into account the patient's estimated GFR. Owing to the heterogeneity of PTH immunoassays currently in use, it is important that laboratories use assay-specific data for reference ranges and the ranges of values to be expected in hyperparathyroidism, hypercalcaemia of malignancy, hypoparathyroidism and renal disease. Failure to do so may result in inappropriate management of patients with renal disease.

The action of PTH on the kidney causes cAMP to appear in the urine. 'Nephrogenous' cAMP has been used in the past as a surrogate measure of PTH. Its current use is limited to the evaluation of PTH resistance (see [Appendix 6.3](#), below).

Classification of hyperparathyroidism. Increases in plasma concentrations of PTH are seen in a variety of circumstances. It is useful clinically to distinguish conditions in which the increased secretion of PTH is a normal physiological response to hypocalcaemia (secondary hyperparathyroidism) from those in which hypersecretion of PTH is the principal abnormality (primary hyperparathyroidism). If secondary hyperparathyroidism persists over long periods then the parathyroids can become hyperplastic. A gradual rise in plasma calcium accompanies this change and overt hypercalcaemia can develop. This is known as tertiary hyperparathyroidism, and arises chiefly in long-term dialysis patients, but it has also been described in patients with malabsorption syndromes. This classification is unsatisfactory in some respects. For example, the kidney transplant recipient with hypercalcaemia due to hyperparathyroidism that developed while on dialysis is difficult to categorize.

An alternative classification, which has the attraction of being more closely related to treatment stratagems, is to describe hyperparathyroidism in relation to the prevailing plasma calcium concentration. Thus, hyperparathyroidism can be described as hypocalcaemic, normocalcaemic or hypercalcaemic.

Vitamin D

Synthesis and metabolism. Vitamins are defined as essential organic compounds that the body cannot synthesize, and are therefore a vital component of the diet. The term 'vitamin D' is a misnomer as the majority of vitamin D is synthesized in the skin by the action of ultraviolet light on 7-dehydrocholesterol ([Fig. 6.3](#)). This produces cholecalciferol (vitamin D₃). Vitamin D is present in a variety of foodstuffs. It can be added to food in the form of ergocalciferol (vitamin D₂ – an additive to margarine and some breakfast cereals and, in the USA, to milk) or may occur naturally in animal products in the form of cholecalciferol. Fish oils are the richest dietary source of cholecalciferol.

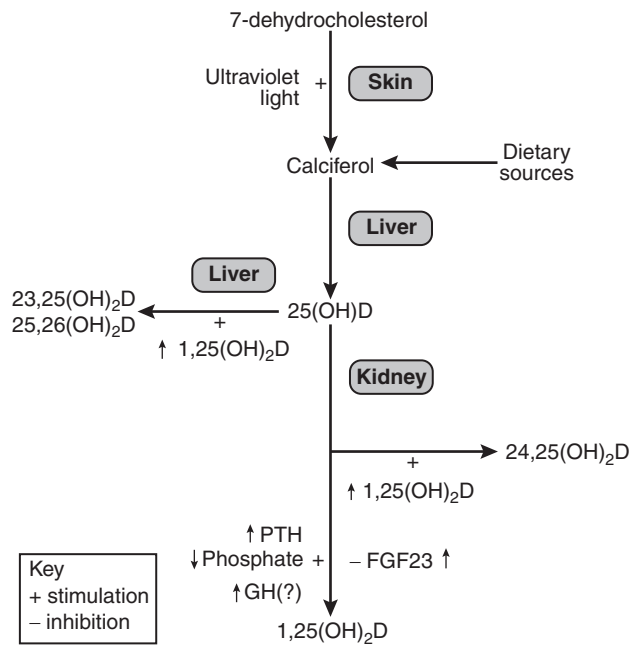


FIGURE 6.3 ■ A summary of vitamin D metabolism. Factors stimulating (+) or inhibiting (–) metabolic conversions are indicated. All the dihydroxylated metabolites of renal or hepatic origin are further metabolized before excretion. Ultimately, most vitamin D is excreted through the bile into the faeces, in the form of inert metabolites. FGF23, fibroblast growth factor 23; GH, growth hormone.

In most respects, these two forms of vitamin D have similar metabolism and action. The calciferols are fat soluble and large quantities can be stored in adipose tissue. They are virtually without biological activity unless they have been hydroxylated. Hydroxylation by the enzyme vitamin D 25-hydroxylase (CYP2R1), takes place in the liver producing 25-hydroxyvitamin D (25(OH)D, calcidiol), which is the principal circulating metabolite and the prohormone for 1,25(OH)₂D (calcitriol). The dependence on dermal synthesis for provision of vitamin D is well illustrated by the marked variation in plasma 25(OH)D concentrations through the year in non-tropical areas of the world, where values are significantly higher in summer and autumn than in winter and spring. The vitamin D metabolites circulate in plasma bound to an α-globulin, vitamin D-binding protein and, to a lesser extent, to albumin (Table 6.2).

Most of the circulating 25(OH)D is metabolized in the liver via intermediates such as 23,25(OH)₂D and 25,26(OH)₂D into inactive metabolites that are excreted into the bile. A small proportion of circulating 25(OH)D undergoes further hydroxylation in the cells of the

proximal renal tubules to produce either 1,25(OH)₂D or 24,25(OH)₂D (secalciferol). The principal active metabolite of vitamin D is 1,25(OH)₂D. Although circulating concentrations of 24,25(OH)₂D are normally ten times greater than those of 1,25(OH)₂D, it has no definite physiological role.

In contrast to hepatic 25-hydroxylation, renal 1α-hydroxylation is closely regulated, being enhanced by low circulating concentrations of phosphate and 1,25(OH)₂D, and by high concentrations of PTH, and suppressed by high concentrations of the bone-derived hormone fibroblast growth factor 23 (FGF23). When the plasma concentration of phosphate is high and PTH low, 1α-hydroxylase activity (enzyme CYP27B1) decreases and that of 24-hydroxylase (enzyme CYP24B) becomes greater. Other factors may also influence renal hydroxylation of vitamin D. In vitro, low calcium can stimulate 1α-hydroxylation independently of PTH, although whether this is significant in vivo is not known. Direct or indirect effects of growth hormone, prolactin, oestrogens and calcitonin have also been described.

1,25-Dihydroxyvitamin D is inactivated by the formation of more polar metabolites, which are excreted primarily into bile. Over 96% of all vitamin D is ultimately eliminated through the bile into the faeces. Since this excretion is largely in the form of inactive metabolites, it is doubtful if there is any enterohepatic recycling of physiological importance.

Although the kidneys are undoubtedly the major source of production of 1,25(OH)₂D, there is evidence from anephric haemodialysis patients that small quantities can be synthesized extrarenally. In pregnancy, 1,25(OH)₂D is synthesized in the placenta, and in a variety of granulomatous diseases (sarcoidosis, tuberculosis, berylliosis) the macrophages of the granulomata can also 1α-hydroxylate vitamin D. In these circumstances, the synthesis is substrate dependent, meaning that the more 25(OH)D that is available, the more 1,25(OH)₂D is synthesized.

Actions. Vitamin D metabolites act via a cytosolic receptor that is translocated to the nucleus where it regulates gene expression. The vitamin D receptor is similar in structure to the steroid and thyroid hormone receptors, having an amino-terminal domain, a central DNA-binding domain with ‘zinc fingers’ and a carboxy-terminal hormone-binding domain. 1,25-Dihydroxyvitamin D has an affinity for the vitamin D receptor that is approximately 1000 times greater than that of 25(OH)D, which in turn is an order of magnitude more potent than 24,25(OH)₂D. However, the circulating concentration of 25(OH)D is substantially greater than that

TABLE 6.2 Circulating concentrations of vitamin D metabolites

Metabolite	Concentration	% Bound to DBP ^a	% Bound to albumin	% Free
Calciferol	2.5–5.0 nmol/L			
25-hydroxyvitamin D	50–125 nmol/L ^a	88	12	0.03%
1,25-hydroxyvitamin D	50–150 pmol/L ^b	85	15	0.5%
24,25-hydroxyvitamin D	1.3–3.8 nmol/L			

^aHigher in summer/autumn than winter/spring; ^bhigher in pregnancy and during puberty.

^cVitamin D binding protein. Interconversions: 10 nmol ≡ 4 μg ≡ 160 international units.

of $1,25(\text{OH})_2\text{D}$, so it is possible that this metabolite has some biologically significant effect.

The principal site of action of $1,25(\text{OH})_2\text{D}$ is the intestine, where it stimulates the production of the calcium channels TRPV6 and TRPV5, and the calcium-binding protein CaBP-9k. In bone, it is a potent osteolytic factor *in vitro*, acting through osteoblast RANK-ligand production to bring about fusion of osteoclast precursors to form mature osteoclasts. The physiological importance of its bone resorbing effect is uncertain. $1,25$ -Dihydroxyvitamin D also has direct effects on the liver and kidney. In the liver, it enhances the production of inactive $25(\text{OH})\text{D}$ derived metabolites that are excreted in the bile, and in the kidney it stimulates the $24,25$ -hydroxylase – both these mechanisms protect against chronically high $1,25(\text{OH})_2\text{D}$ concentrations.

Several cells and tissues other than those involved in calcium and phosphorus metabolism, including macrophages, activated T-lymphocytes, the placenta and keratinocytes, express vitamin D receptors and/or the 1α -hydroxylase enzyme, suggesting that their functions might be regulated by components of the vitamin D system. Many malignant cells also express the vitamin D receptor. There is evidence that $1,25(\text{OH})_2\text{D}$ decreases the growth of several cancer cell lines, and slows or prevents the growth and/or development of cancers and autoimmune diseases in rodents. Genetic ablation of vitamin D signaling, by deleting the genes encoding either the vitamin D receptor or 1α -hydroxylase, leads to renin-dependent hypertension and an increased propensity to tumour development. The consequences of complete absence of vitamin D signaling in rodents, however, are likely to be very different from minor decreases in $25(\text{OH})\text{D}$ in humans. At present, there is no robust clinical trial evidence that vitamin D supplements or bioactive vitamin D analogues favourably influence important outcomes such as cancer, hypertension or autoimmune diseases. Accordingly, vitamin D supplements are not indicated in the management of non-skeletal diseases.

Synthetic vitamin D analogues. A variety of $1,25(\text{OH})_2\text{D}$ analogues have been synthesized. 1α -Hydroxyvitamin D_3 (alfacalcidol) is rapidly metabolized by the liver into $1,25(\text{OH})_2\text{D}$ and, although having roughly half the potency of $1,25(\text{OH})_2\text{D}$ on a weight basis, is similarly useful in conditions where the 1α -hydroxylation step is impaired. A similar preparation, 1α -hydroxyvitamin D_2 (doxercalciferol) is also marketed. Other analogues such as $26,27$ -hexafluorocalcitril are more potent calcaemic agents.

A number of analogues with attenuated calcaemic properties have also been developed. Examples include calcipotriol (used in the treatment of psoriasis), paricalcitol (used in some countries in the treatment of hyperparathyroidism in chronic renal failure) and eldecalcitol (which has been investigated in osteoporosis). The divergent effects of these analogues on cell differentiation, PTH secretion and calcium homeostasis probably reflect pharmacokinetic differences rather than distinct $1,25(\text{OH})_2\text{D}$ -sensitive pathways, since there is no evidence that they act through different receptors.

Measurement of vitamin D metabolites. Measurements of $25(\text{OH})\text{D}$ and $1,25(\text{OH})_2\text{D}$ can be of value in clinical practice, though the indications for $1,25(\text{OH})_2\text{D}$ measurement are relatively few (e.g. obscure causes of hypercalcaemia). 25 -Hydroxyvitamin D is the best measure of vitamin D status. While its synthesis is dependent upon 25 -hydroxylation in the liver, even advanced liver disease does not seem to be a limiting factor in its production. Free concentrations are normal in most patients with cirrhosis, though total concentrations may be low because of associated hypoproteinaemia. $1,25$ -Dihydroxyvitamin D is present in concentrations only 1000th of those of its precursor (see Table 6.2). Its production is closely regulated and so is not tied to concentrations of its precursor, except in patients with severe vitamin D deficiency or in granulomatous disease. Renal disease has a major impact on its circulating concentrations, as do states of parathyroid over- or underactivity.

A variety of assays have been used in the past, and a number have been withdrawn from the market because of unsatisfactory performance characteristics. Cross-calibration of available assays has not always been within acceptable limits. The reference method uses liquid chromatography with tandem mass spectrometry (MS/MS), but as this is an expensive assay to run, many laboratories use immunoassays based on automated platforms. The acceptability of this compromise depends on the performance characteristics and calibration of the particular assay platform. Some of the commercially available assays have limited ability to measure $25(\text{OH})\text{D}_2$, which may be important when supplementation has been provided in the form of vitamin D_2 . It is evident that significant variation can occur in measurement of vitamin D metabolites, as a result both of differences between the available assays and in the implementation within individual laboratories.

Calcitonin

Calcitonin is a 32-amino acid peptide produced in humans in the parafollicular cells (C cells) of the thyroid gland. It contains a disulphide bridge and a proline-amide group at the C-terminus. Both of these and the full amino acid sequence are necessary for biological activity. It is metabolized in the kidneys and has a plasma half-life of approximately 5 min.

Secretion of calcitonin is stimulated by an increase in plasma calcium concentration and it is also released in response to a number of gut hormones (e.g. gastrin, glucagon, secretin, cholecystokinin-pancreozymin). It reduces plasma calcium concentration by a direct effect on osteoclasts, which have a cell surface calcitonin receptor linked to adenylate cyclase. It causes contraction of osteoclasts and an acute reduction in osteoclastic bone resorption. It also acts on the kidneys, where it decreases renal tubular reabsorption of calcium and phosphate.

Despite these clearly documented actions, its physiological significance is uncertain because its actions appear to be transient. Furthermore, chronic calcitonin deficiency (as in patients post-thyroidectomy) or excess (as in medullary carcinoma of the thyroid) do not result in significant changes in bone or mineral metabolism.

There are a number of different forms of calcitonin in the circulation, many of which are not biologically active. Thus, different immunoassays may detect different molecular species. The measurement of calcitonin does not form a significant part of the clinical evaluation of calcium metabolism, but it is an important tumour marker for medullary carcinoma of the thyroid.

Procalcitonin. Procalcitonin, the 116 amino acid peptide precursor of calcitonin, is produced not only by thyroid C-cells but also by the neuroendocrine cells of the lung and the intestine. The amount of procalcitonin produced by the cells of the lung and the intestine increases in response to proinflammatory stimuli, especially of bacterial origin. Measurement of procalcitonin can be used as a marker of severe sepsis and generally correlates well with the degree of sepsis.

Other hormones

A number of hormones influence calcium and bone metabolism *in vitro*, in normal physiology or in specific pathological circumstances. The oestrogen deficiency that develops at the time of the menopause results in increases in cytokine production which leads to increases in bone resorption. Direct effects of oestrogen on osteoclasts and on osteoblastic production of RANKL and osteoprotegerin are also important. Calcium absorption in the gut and reabsorption in the kidney are both reduced, and plasma calcium concentrations rise. These changes lead to a reduction in bone mass, but are reversible with oestrogen replacement therapy. Similarly, testosterone deficiency in men predisposes to osteoporosis, partly through loss of its direct anabolic effects on muscle and on osteoblasts, and partly through diminished conversion to oestrogen. Glucocorticoids can cause marked osteoporosis. These hormones act at a number of sites and reduce osteoblastic activity, induce apoptosis of osteoblasts and osteocytes, reduce intestinal calcium absorption, increase urinary calcium loss and probably increase bone resorption. In children, glucocorticoid excess causes marked growth retardation. Glucocorticoid deficiency may be associated with hypercalcaemia.

Growth hormone accelerates linear growth in children and increases plasma phosphate concentration by increasing renal tubular reabsorption of phosphate. These effects are probably mediated by growth hormone-induced changes in the production of insulin-like growth factor 1 (IGF-1, somatomedin C). There is controversy as to the effect of the growth hormone–somatomedin axis on bone density.

Subjects with high total body fat mass have been shown to have higher bone densities. This appears to be mediated by a number of hormonal mechanisms. Insulin is anabolic to bone *in vitro*, and obesity is usually associated with hyperinsulinaemia. At least two other anabolic peptides, amylin and preptin, are co-secreted with insulin and may contribute to the relationship. Leptin has been shown *in vitro* to stimulate osteoblast growth directly and to inhibit osteoclast development, so it is also likely to be an important factor. In contrast, when leptin is administered into the third ventricle of the brains of

experimental animals, there is a profound reduction in appetite and fat mass, with resultant loss of bone mass. There is some evidence that central leptin administration activates the sympathetic nervous system and that this might also contribute to bone loss.

Thyroid hormones are potent stimulators of bone resorption, and thyrotoxicosis can be associated with decreased bone density and hypercalcaemia. Prostaglandins of the E series also stimulate bone resorption *in vitro*. Parathyroid hormone-related peptide (PTHrP) is a 141-amino acid peptide that is responsible for a large proportion of cases of humoral hypercalcaemia of malignancy. Eight of the 13 amino-terminal amino acids are identical to those found in PTH itself, and it appears to act on the PTH receptor. There is evidence for its production in both the fetal parathyroids and in the lactating breast, raising the possibility that it has a physiological role in these contexts. A carboxy-terminal fragment of PTHrP may act as an inhibitor of osteoclastic bone resorption. Direct measurement of PTHrP by immunoassay is possible, but is rarely necessary in clinical practice, since the malignant disease underlying humoral hypercalcaemia is frequently evident at the time that the elevated plasma calcium is recognized.

Biochemical assessment of calcium metabolism

In addition to the assays of calcitropic hormones discussed above, estimation of calcium concentrations in the circulation and of calcium fluxes in the gut and kidney are of value in assessing patients with disorders of calcium metabolism.

Plasma calcium

As indicated previously, total plasma calcium comprises protein-bound, complexed and ionized fractions. Thus changes in plasma protein concentrations or acid–base status will affect the relationship between ionized and total calcium. Total serum calcium remains the most commonly measured index. It should always be accompanied by a measurement of the serum albumin concentration from which a ‘corrected serum calcium’ can be calculated:

$$\begin{aligned} \text{corrected serum concentration (mmol/L)} = \\ \text{measured calcium concentration} + \\ 0.02(40 - \text{albumin concentration (g/L)}) \end{aligned}$$

Laboratories should derive their own correction formulae based on their own particular calcium and albumin assays, as both the mean normal albumin concentration (assumed here to be 40 g/L) and the amount of calcium bound to each gram of albumin (assumed here to be 0.02 mmol/g) can vary. The corrected calcium is only an approximation, and when there are significant dysproteinaemias or acid–base disturbances, ionized calcium should be measured directly. This measurement is also useful in patients to whom large quantities of citrated blood products have been administered, since these individuals have an increased proportion of plasma calcium complexed to citrate. Ionized calcium is

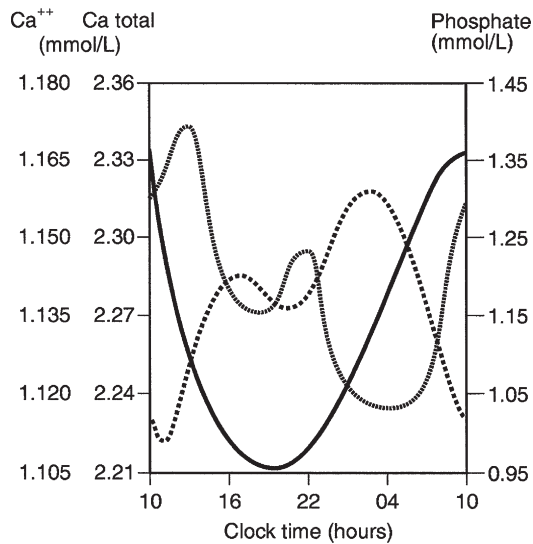


FIGURE 6.4 ■ Diurnal variation in total plasma calcium (dotted line), ionized calcium (solid line) and phosphate (broken line) in healthy subjects. (From Markowitz M et al. 1981 *Science* 213:672–674, with permission from the American Association for the Advancement of Science).

measured using an ion-selective electrode. Specimens must be obtained anaerobically to minimize changes in specimen pH. The pH is also usually measured and results adjusted to a pH of 7.4.

There are diurnal variations in both total and ionized plasma calcium concentrations (Fig. 6.4). Significant changes in plasma calcium take place following the ingestion of a calcium-rich meal or of calcium supplements. Posture and the application of tourniquets, both of which can influence plasma protein concentrations, may also influence total calcium estimations. Total plasma calcium rises significantly (by approximately 0.1 mmol/L) at the menopause.

Intestinal calcium absorption

This is not routinely measured in clinical practice, but knowledge of it is sometimes of value, e.g. in some patients with kidney stones and hypercalciuria. In the past, it was measured by metabolic balance, which required a period of 7–10 days as an inpatient, during which time subjects were given a diet with a constant calcium intake, and calcium absorption was estimated from precise measurements of faecal and urinary calcium output.

Calcium absorption can be measured with a variety of stable or radioisotope techniques in which an oral calcium load with a tracer is administered in the fasting state and absorption efficiency is estimated from the rate of appearance of the isotope in the blood. The size of the oral calcium load accompanying the isotope is important. Low doses reflect better the active duodenal transport, whereas high doses better reflect net intestinal absorption. The increment in urine calcium following the administration of a known calcium load has also been used as an index of calcium absorption (see Appendix 6.1, below). The clinical utility of these tests is limited and they are rarely used nowadays.

Urinary calcium

Calcium can be measured either in a 24 h urine or in urine collected after an overnight fast. In the latter instance, the subject empties the bladder on rising, has a glass of water and then collects the urine sample at any time between 30 min and 2 h later. In both instances, acid (usually 6 M hydrochloric acid) is added either before collection or subsequently, to prevent crystallization of calcium salts. Creatinine is usually measured on both 24 h and fasting urine samples – in the former case to allow assessment of completeness of collection and in the latter as a denominator in relation to which the concentration of calcium (and other urinary constituents) can be expressed.

The reference range for 24 h urine calcium is 1–7.5 mmol/24 h in men and 1–6.25 mmol/24 h in women. In the steady state, this reflects the absolute dietary intake of calcium and the net proportion absorbed from the intestine. Urine calcium excretion is usually increased with hypercalcaemia of any cause, and its measurement contributes little to the differential diagnosis of hypercalcaemia, except in the diagnosis of familial hypocalciuric hypercalcaemia.

Fasting urine calcium can be expressed as a simple molar ratio to creatinine (reference values 0.10–0.30 in adults). This is considered to be an index of bone resorption, but newer markers of bone resorption are preferred in clinical practice (see Chapter 31). If the calcium:creatinine ratio is multiplied by the plasma creatinine concentration, the resulting product is the calcium excretion per litre of glomerular filtrate (Ca_E). This is useful in assessing the contribution of abnormal renal tubular handling of calcium to disorders of plasma calcium homeostasis (see Appendix 6.2, below).

Indices of bone turnover

Disorders of plasma calcium are often accompanied by abnormalities of bone turnover, so assessment of bone turnover can be useful. These measures are discussed in more detail in Chapter 31.

Hypercalcaemia

Hypercalcaemia develops when the rate of entry of calcium into the ECF from bone and gut exceeds the capacity of the kidney to excrete it. Diminished excretory capacity can be the result of either increased renal tubular reabsorption (an important component, for example, in the hypercalcaemia of primary hyperparathyroidism and familial hypocalciuric hypercalcaemia) or of a reduction in glomerular filtration rate (in sarcoidosis, for example, hypercalcaemia is more frequently encountered in subjects with impaired renal function). However, hypercalcaemia itself can affect these processes, thus initiating a vicious cycle with worsening hypercalcaemia. By its actions on the kidney, hypercalcaemia leads to polyuria, volume depletion and thus to a reduction in glomerular filtration rate and enhanced proximal tubular reabsorption of sodium, which carries calcium with it. This situation, termed ‘dis-equilibrium hypercalcaemia’, is a medical emergency. It is important both to recognize this and to differentiate it from ‘equilibrium hypercalcaemia’ (e.g. due to mild

primary hyperparathyroidism or familial benign hypercalcaemia), where the plasma calcium remains stable.

Whether hypercalcaemia causes symptoms depends both upon the degree of elevation of plasma calcium and the rate at which it has risen. When plasma calcium is <3.0 mmol/L, a large proportion of patients are asymptomatic. The clinical features of severe hypercalcaemia are summarized in [Box 6.1](#).

Causes of hypercalcaemia

[Box 6.2](#) lists the most frequently encountered forms of hypercalcaemia.

Primary hyperparathyroidism. Primary hyperparathyroidism is the most common cause of hypercalcaemia presenting outside hospital. Prior to the advent of large automated analysers, it was regarded as a rare disease, but is now a common incidental finding. Its prevalence is approximately 1 per 1000 and it is most frequently diagnosed in the 6th decade of life, when it is two or three times as common in women as in men. To some extent, this can be explained by the postmenopausal rise in plasma calcium concentration. The normal distribution curve is shifted to the right and, therefore, a higher proportion of postmenopausal women must necessarily have plasma calcium concentrations above the upper limit of normal for younger people. In younger patients, the sex incidences are equal.

In 90% of patients, primary hyperparathyroidism is attributable to a single parathyroid adenoma, and most of the remaining cases are attributable to four gland hyperplasia, sometimes as a part of the multiple endocrine neoplasia (MEN) syndrome type 1 (see Chapter 41). This syndrome accounts for the majority of patients with familial hyperparathyroidism. Parathyroid carcinoma, which

BOX 6.2 Differential diagnosis of most frequently encountered forms of hypercalcaemia

Most common

Primary hyperparathyroidism

- Sporadic
- Familial (see [Table 6.3](#))

Malignant disease

- Humoral hypercalcaemia of malignancy (PTHrP)
- Widespread skeletal metastases (most commonly breast cancer)
- Haematological malignancy (multiple myeloma, adult T cell leukaemia/lymphoma)^a

Less common

- Granulomatous disease^a (e.g. sarcoidosis, tuberculosis, histoplasmosis, leprosy)
- Toxicity with vitamin D or its derivatives
- Persistent hyperparathyroidism after renal transplantation
- Severe thyrotoxicosis

Rarer causes of hypercalcaemia are summarized in [Table 6.4](#)

^aHypercalcaemia is typically steroid-responsive.

usually presents with severe hypercalcaemia and very high PTH concentrations, accounts for $<1\%$ of all cases.

Adenomatous primary hyperparathyroidism appears to have a clonal origin. Two genetic mechanisms have been elucidated ([Table 6.3](#)). The PTH gene is located on the opposite side of the centromere of chromosome 11 from the gene (*PRAD1*) that encodes cyclin D1, a proto-oncogene that regulates cell growth. In about 20% of sporadic parathyroid tumours, it has been shown that while one chromosome is intact, the other copy has undergone centromeric inversion, a rearrangement that places the PTH regulatory elements under the influence of cyclin D1, allowing deregulated parathyroid gland growth.

Germline mutations underlie several of the inherited syndromes that include primary hyperparathyroidism. The mutated gene in multiple endocrine neoplasia 1 (*MEN1*) is also located on chromosome 11, and encodes a protein, menin, that acts as a tumour suppressor. Rare instances have been reported of kindreds with germline *MEN1* mutations in which primary hyperparathyroidism is the only clinical feature. Primary hyperparathyroidism is less frequently (~20%) a feature of the MEN 2 syndromes, which result from activating mutations of the gene encoding the RET tyrosine kinase receptor. Germline mutations in the tumour suppressor gene *CDC73*, which encodes the protein parafibromin, causes the inherited disorder, hyperparathyroidism-jaw tumour syndrome. Somatic mutations in this gene are also present in a high proportion of sporadic parathyroid carcinomas.

Primary hyperparathyroidism may present with the symptoms of hypercalcaemia outlined in [Box 6.1](#), but frequently patients are symptom-free. The prevalence of osteoporosis is probably increased in patients with primary hyperparathyroidism, especially at skeletal sites enriched for cortical bone. The classic skeletal manifestations of

BOX 6.1 Clinical features and associations of hypercalcaemia

(Typically only present if hypercalcaemia is ≥ 3 mmol/L)

Neuropsychiatric

- Lethargy, depression
- Confusion, coma
- Hypotonia, hyporeflexia, muscle weakness

Gastrointestinal

- Anorexia, nausea, vomiting
- Abdominal pain
- Constipation
- Pancreatitis

Cardiovascular

- Electrocardiographic changes (reduced QT interval, prolonged PR interval)
- Arrhythmias

Renal

- Polyuria, polydipsia, volume depletion
- Reduced glomerular filtration rate
- Calculi, nephrocalcinosis

TABLE 6.3 Molecular genetics of primary hyperparathyroidism

Disorder	Inheritance OMIM number	Gene	Somatic/ germline	Protein product	Pathogenesis
Parathyroid adenoma	Sporadic	<i>CCND1</i>	somatic	cyclin D1	Oncogenic activation
Familial isolated hyperparathyroidism	AD 145000	<i>MEN1</i>	germline	menin	Tumour suppressor inactivation
Multiple endocrine neoplasia 1	AD 131100	<i>MEN1</i>	germline	menin	Tumour suppressor inactivation
Multiple endocrine neoplasia 2a	AD 171400	<i>RET</i>	germline	ret	Oncogenic activation
Hyperparathyroid-jaw tumour syndrome	AD 145000	<i>CDC73</i>	germline	parafibromin	Tumour suppressor inactivation
Parathyroid carcinoma	Sporadic 608266	<i>CDC73</i>	somatic	parafibromin	Tumour suppressor activation
Familial hypocalciuric hypercalcaemia type 1	AD 145980	<i>CASR</i>	germline	CaSR	Decreased set-point for PTH secretion
type 2	AD 145981	<i>GNA11</i>	germline	G α_{11}	
type 3	AD 600740	<i>AP2S1</i>	germline	AP17	
Neonatal severe hyperparathyroidism	AR 239200	<i>CASR</i> (<i>bi-allelic</i>)	germline	CaSR	Parathyroid cell proliferation

AD, autosomal dominant; AR, autosomal recessive; CaSR, calcium sensing receptor

primary hyperparathyroidism, which include subperiosteal phalangeal resorption, a 'salt and pepper' appearance in the skull and bone cysts ('brown tumours') of the long bones or jaw (collectively termed 'osteitis fibrosa cystica'), are features of severe, longstanding disease, and are now uncommon in the developed world. Approximately 15% of patients have renal complications in the form of nephrolithiasis or nephrocalcinosis.

Familial hypocalciuric hypercalcaemia. Familial hypocalciuric hypercalcaemia (FHH) (also known as benign familial hypercalcaemia) is an autosomal dominant condition in which mild hypercalcaemia with relative hypocalciuria is present throughout life. Patients with this disorder are most frequently heterozygous for inactivating mutations in the calcium sensing receptor (CaSR). The 'set-point' for PTH release is raised, and mild PTH-dependent hypercalcaemia results. The CaSR is also expressed in the renal tubules, where it regulates renal tubular calcium reabsorption in a PTH-independent fashion. The inactivating *CASR* gene mutation promotes inappropriately avid tubular calcium reabsorption in the face of hypercalcaemia, explaining the relative hypocalciuria. Patients with FHH are usually asymptomatic, although non-specific symptoms of lethargy and polydipsia have been observed in some instances. There may be an increased incidence of pancreatitis, but the frequency of nephrolithiasis and peptic ulcer disease is the same as in the general population. In most cases, the condition follows a benign course and the greatest danger to patients is being misdiagnosed as having adenomatous primary hyperparathyroidism and thus having inappropriate parathyroidectomy. Familial hypocalciuric hypercalcaemia accounts for about 2% of asymptomatic hypercalcaemia.

Neonatal severe primary hyperparathyroidism occurs in newborns who have inherited two inactivating *CASR* mutations – one from each parent. It presents as severe, life-threatening hypercalcaemia. Urgent total parathyroidectomy is indicated.

Hypercalcaemia of malignancy. Malignancy can result in hypercalcaemia by four general mechanisms:

- by the secretion into the circulation of factors that increase bone resorption and/or decrease urine calcium loss (humoral hypercalcaemia of malignancy, HHM)
- as a result of osteolytic metastases that lead to local bone destruction
- osteolysis from bone marrow infiltration in haematological disorders
- more rarely, accelerated conversion of plasma 25(OH)D to 1,25(OH)₂D by some haematological malignancies, leading to vitamin D-dependent hypercalcaemia.

Parathyroid hormone-related peptide is the main mediator of humoral hypercalcaemia of malignancy. It is synthesized and secreted by a number of tumours, particularly those of epithelial cell origin (e.g. squamous cell carcinoma of the bronchus, renal cell carcinoma, breast carcinoma) and produces biochemical changes similar to those of primary hyperparathyroidism, although the degree of hypercalcaemia is often more severe in HHM, and plasma 1,25(OH)₂D concentrations tend to be normal or low rather than increased.

Hypercalcaemia associated with osteolytic metastases is most commonly seen in advanced breast cancer, but can occur with other solid tumours that spread to bone. Hypercalcaemia is very common in some haematological malignancies, particularly multiple myeloma and acute human T cell leukaemia-lymphoma virus (HTLV)-1-associated disease. The bone destruction is mediated by

osteoclasts, activated by factors released by the tumour. A number of different cytokines have been implicated, including RANKL and TNF- α . In addition, bone formation may be impaired at sites of myelomatous skeletal disease, as a result of inhibition of Wnt signalling through the lipoprotein receptor-related protein 5 (LRP5). In contrast to the hypercalcaemia associated with metastases from solid tumours, hypercalcaemia associated with haematological malignancy typically responds rapidly to glucocorticoid treatment. In the rare cases of patients with hypercalcaemia associated with haematological malignancies that elaborate 1 α -hydroxylase, the hypercalcaemia also responds rapidly to glucocorticoid treatment.

Granulomatous disease. The macrophages in granulomata occurring in sarcoid tissue, pulmonary tuberculosis and berylliosis are capable of 1 α -hydroxylating 25(OH)D, independently of normal homeostatic regulation. The production of 1,25(OH)₂D is dependent on disease activity and circulating 25(OH)D concentrations. Thus, during the summer months, patients with active disease have increased intestinal calcium absorption. This is usually evident as hypercalciuria, but some patients, usually those with some pre-existing renal impairment, become hypercalcaemic. 1,25-Dihydroxyvitamin D production can be rapidly suppressed (and hypercalcaemia corrected) by glucocorticoids, chloroquine or ketoconazole.

Vitamin D toxicity. Excessive intake of vitamin D or its analogues can produce hypercalcaemia. This may be iatrogenic or the result of self-medication. On withdrawal of vitamin D, hypercalcaemia resolves. The rate of reversal

is much more rapid in patients intoxicated with alfalcidol or calcitriol than it is in patients intoxicated by calciferol. In the former instances, plasma calcium reverses with a half-time of 1–5 days, but in the latter, the half-time is of the order of 10–30 days. Women being treated with vitamin D or its analogues for hypoparathyroidism may become hypercalcaemic during lactation or after stopping exogenous oestrogen therapy.

The vast majority of adult hypercalcaemic patients (>98%) will have one of the above diagnoses, but there are rarer causes of hypercalcaemia (summarized in Table 6.4). Hypercalcaemia in dialysis and kidney transplant recipients is discussed in Chapter 31. Hypercalcaemia in infants and children is much less common than in adults, and a different range of diagnoses needs to be considered (Box 6.3).

Investigation of hypercalcaemia

Figure 6.5 sets out an approach to the investigation of hypercalcaemia. Having confirmed that true hypercalcaemia exists, a clinical assessment will in many cases, point to the appropriate diagnosis by revealing evidence of underlying diseases or medications that may be contributing. The next stage is biochemical assessment, in which PTH measurement is pivotal. The currently available assays of intact PTH will usually produce either an elevated or suppressed value in hypercalcaemic subjects, thus establishing or excluding the parathyroids as the cause of the hypercalcaemia.

A normal intact PTH concentration in the face of an elevated plasma calcium may be consistent with either primary hyperparathyroidism or with FHH. The ratio of calcium clearance to creatinine clearance in a 2 h fasting

TABLE 6.4 Uncommon causes of hypercalcaemia

Cause	Probable mechanisms	Other biochemical features (plasma)	Treatment	Comments
Thiazides	↑ Renal tubular reabsorption	–	Withdrawal	Often coexisting primary hyperparathyroidism
Lithium	? Enhanced sensitivity to PTH or ↑ PTH secretion	–	Withdrawal	Often coexisting primary hyperparathyroidism
Oestrogens/tamoxifen	↑ Bone resorption	–	Transient hypercalcaemia only	In women with skeletal metastases from breast cancer
Vitamin A toxicity	↑ Bone resorption	–	Withdrawal	Also described with retinoic acid derivatives
Immobilization	↑ Bone resorption	↓ [PTH], ↑ [phosphate]	Mobilize or inhibitors of bone resorption	Occurs in individuals with high bone turnover rate (e.g. tetraplegia, extensive Paget disease)
Acute kidney injury (AKI) – diuretic phase	↑ Calcium absorption	↑ [1,25(OH) ₂ D]	Resolves spontaneously	Occurs in recovery phase of AKI induced by rhabdomyolysis
Multiple organ dysfunction/systemic inflammatory response syndrome	Unknown	? ↑ [1,25(OH) ₂ D]	Resolves spontaneously	Poor response to bisphosphonates
Islet cell tumours/phaeochromocytoma	? ↑ Bone resorption	? ↑ [PTHrP]	Surgery or inhibitors of bone resorption	–
Addison disease	↑ Renal tubular reabsorption	↓ [Na ⁺], ↑ [K ⁺]	Fluids, steroid replacement	–
Milk-alkali syndrome	↑ Calcium intake	–	Withdrawal	Often coexisting primary hyperparathyroidism

BOX 6.3 Causes of hypercalcaemia in infants and children**Hyperparathyroidism**

- Severe neonatal hyperparathyroidism (bi-allelic *CASR* mutations)
- Multiple endocrine neoplasia type 1
- Parathyroid carcinoma

Malignancy

- Acute leukaemia
- Solid tumours

Syndromic

- Williams–Beuren syndrome (microdeletion of contiguous genes on chromosome 7q11.23)
- Jansen syndrome (mutation of PTH/PTHrP receptor)

Vitamin D-related

- Iatrogenic
- Infantile hypercalcaemia (inactivating *CYP24B* mutations)
- Granulomatous disease (tuberculosis, sarcoidosis)

Immobilization**Miscellaneous**

- Hypophosphatasia (*TNSALP* mutations)
- Subcutaneous fat necrosis
- Congenital lactase deficiency
- Primary hyperoxaluria type 1

urine collection is useful in distinguishing between these two possibilities – in FHH the ratio is usually < 0.01. However, clear distinction between adenomatous primary hyperparathyroidism and FHH is not always possible on biochemical grounds. Studying family members and/or sequencing the *CASR* gene may be necessary.

If plasma PTH concentration is low, then a variety of other investigations can be pursued. The order in which these are undertaken will be determined by the clinical clues. Cancer-associated hypercalcaemia is almost always a late complication of the disease, which is therefore either known or readily diagnosed when the hypercalcaemia is recognized. A search for malignancy may include radioisotope bone scanning, chest and abdominal imaging and serum and urine protein electrophoresis. Assays for PTHrP are now available, but are rarely needed in the evaluation and treatment of hypercalcaemia. Plasma PTHrP concentrations are elevated in most patients with humoral hypercalcaemia of malignancy, and are low in healthy subjects (<2 pmol/L). Parathyroid hormone-related peptide concentrations are also increased in about half of women with lytic bone metastases from breast cancer, irrespective of whether they are hypercalcaemic. In difficult cases, measurement of 1,25(OH)₂D may be of value. Elevated concentrations are seen in granulomatous disease, some lymphomas and sometimes in primary hyperparathyroidism, and subnormal concentrations are usually found in HHM. Measurement of 25(OH)D will allow confirmation of vitamin D intoxication, but not if it is caused by 1 α -hydroxylated compounds such as alfacalcidol, calcitriol, paricalcitol or calcipotriol.

A number of investigations previously used in the differential diagnosis of hypercalcaemia have limited utility. Hypophosphataemia is a feature of both primary hyperparathyroidism and HHM, and phosphate concentrations tend to be increased in sarcoidosis and vitamin D toxicity or when there is renal impairment. Measurement of plasma chloride has been used because of the mild metabolic acidosis associated with primary hyperparathyroidism. While chloride concentrations do tend to be higher in primary hyperparathyroidism than in HHM, this measurement has very little clinical utility in individual patients. The same considerations apply to plasma magnesium concentrations. Although plasma magnesium is higher in FHH than in primary hyperparathyroidism, it is of limited diagnostic value. Since both PTH and PTHrP stimulate renal adenylate cyclase, measurement of cyclic adenosine monophosphate excretion has no discriminant value.

Skeletal radiology may be of help, showing myeloma deposits (that may be missed by bone scintigraphy) or showing evidence of other malignancy or revealing the changes of hyperparathyroidism. Occasionally, use is still made of the glucocorticoid suppression test (prednisolone 30 mg/day for a period of ten days). In primary hyperparathyroidism, no change in plasma calcium occurs, but in granulomatous disease, vitamin D intoxication and in haematological malignancies, a significant fall in plasma calcium usually occurs. The analysis of renal tubular calcium handling (see [Appendix 6.2](#), below) can provide information on the mechanisms underlying hypercalcaemia, but does not help to differentiate primary hyperparathyroidism from HHM.

In summary, measurement of intact PTH, accompanied by clinical history and examination and targeted radiological investigations will allow the cause of hypercalcaemia to be ascertained quite quickly in the vast majority of patients.

Treatment of hypercalcaemia

Specific details of the treatment of hypercalcaemia are beyond the scope of this text, but its principles will be reviewed. In patients with mild asymptomatic primary hyperparathyroidism, long-term follow-up has revealed stability of plasma calcium and renal function, and low rates of disease complications. Consequently, some authorities feel that no intervention is necessary in patients who have not had renal stones and are at low fracture risk, although this remains an area of controversy. Either medical treatment with bisphosphonates or surgical correction are effective treatments for improving bone mineral density in patients with primary hyperparathyroidism who are at increased risk of fracture. When definitive treatment is necessary, as in patients who develop urolithiasis or severe hypercalcaemia (>3 mmol/L), removal of the causative adenoma by an experienced surgeon is usually curative. Recurrence is rare, except patients in whom hyperparathyroidism is caused by multiglandular hyperplasia, as in the type 1 multiple endocrine neoplasia syndrome. Calcimimetic drugs, such as cinacalcet, which act as allosteric agonists of the CaSR present on the parathyroid chief cells, suppress PTH secretion and lower plasma calcium in PTH-dependent hypercalcaemia. Studies indicate that treatment with

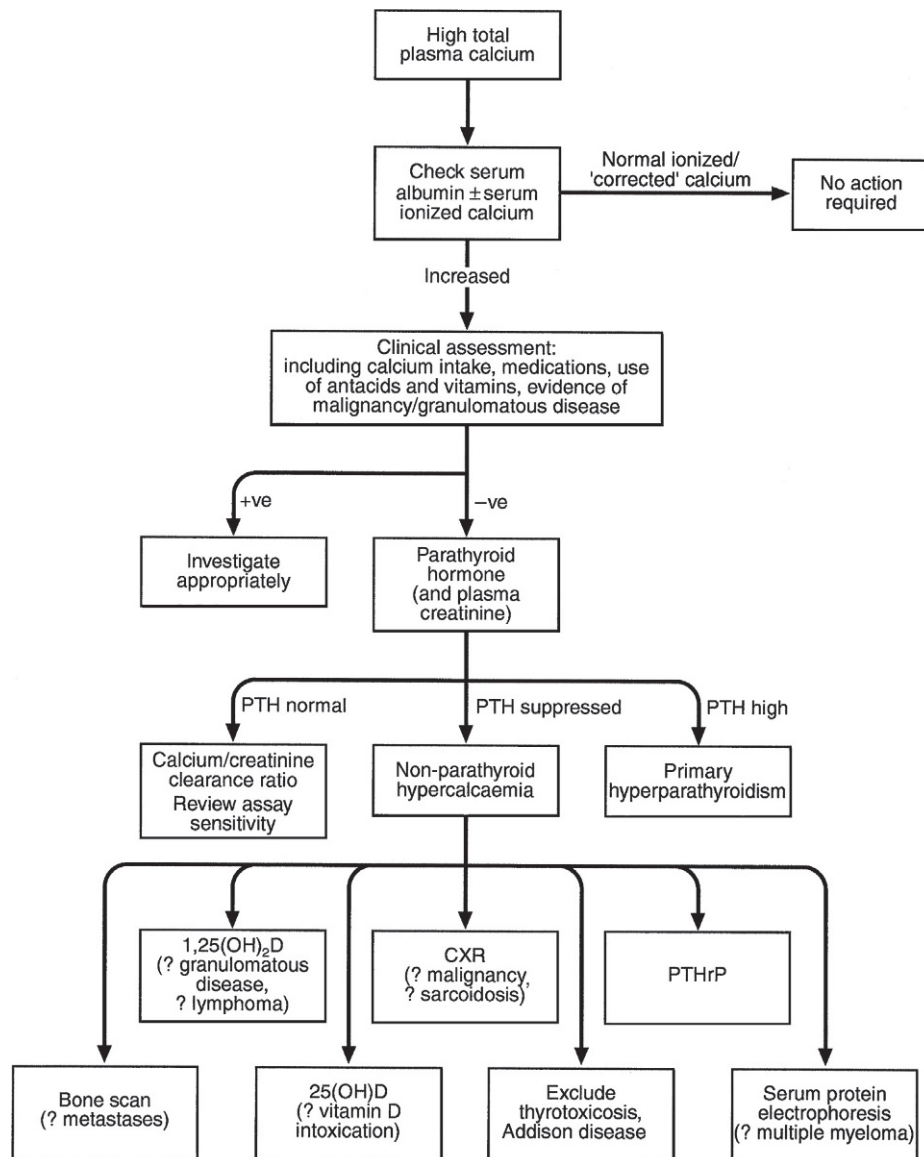


FIGURE 6.5 ■ A flow diagram for investigation of hypercalcaemia.

cinacalcet can maintain normocalcaemia for at least five years in the vast majority of patients with mild primary hyperparathyroidism, and that it is moderately effective in lowering plasma calcium in patients with severe, symptomatic PTH-dependent hypercalcaemia. Thus, the calcimimetics may acquire a role in the future in treating patients with severe, symptomatic primary hyperparathyroidism or parathyroid cancer in whom surgery has either failed or is not possible. In FHH, no therapy is required.

Individuals with disequilibrium hypercalcaemia are almost invariably volume depleted, and the first step in their management is rehydration with intravenous normal saline. This will be associated with increased renal blood flow and a resultant increased clearance of calcium through the kidneys. In those in whom hypercalcaemia is secondary to increased bone resorption, therapy with an inhibitor of bone resorption is appropriate. This is now satisfactorily achieved with intravenous infusions of bisphosphonates (e.g. pamidronate or zoledronate), with the majority of patients gradually returning to normocalcaemia within about a week.

In the past, forced saline diuresis has been used to increase renal clearance of calcium. This is potentially hazardous, particularly in the elderly, in whom congestive cardiac failure and electrolyte imbalances may cause significant morbidity. The infusion of phosphate, which can cause widespread metastatic calcification, is another dangerous and redundant therapy for hypercalcaemia. Steroids are helpful in haematological malignancies and in sarcoidosis.

Hypocalcaemia

Hypocalcaemia is a relatively uncommon finding, but true hypocalcaemia usually indicates a significant underlying abnormality requiring diagnosis and treatment.

Clinical features

Hypocalcaemia results in increased excitability of neuromuscular tissue. This may become clinically manifest as tetany, usually presenting as carpopedal spasm or, in

milder cases, as paraesthesiae in the perioral region or fingers. Laryngeal stridor and seizures may also occur. The electrocardiogram shows prolongation of the QT and ST intervals, and patients may develop arrhythmias, heart block and congestive heart failure.

In longstanding hypocalcaemia, calcification of the basal ganglia, psychiatric disturbances and cataracts may occur. Hypocalcaemic children may show abnormalities in dental development.

Causes of hypocalcaemia

Causes of hypocalcaemia are set out in [Box 6.4](#). Because hypoalbuminaemia is observed in patients with a wide variety of other diagnoses, it is mandatory to make allowance for the plasma albumin concentration or to measure ionized calcium when entertaining the diagnosis of hypocalcaemia. The common causes of chronic hypocalcaemia relate to abnormalities in the synthesis, secretion or action of PTH or $1,25(\text{OH})_2\text{D}$, or both these hormones.

Hypoparathyroidism. The most common type of hypoparathyroidism is that occurring after surgery to the thyroid or, occasionally, to other structures in the neck. It varies widely in severity and, in some patients, is tran-

BOX 6.4 Differential diagnosis of hypocalcaemia

Hypoparathyroidism

- Genetic (see [Table 6.5](#))
- Idiopathic
- Post-surgical, post-neck irradiation
- Severe hypomagnesaemia
- Infiltrative
 - β -thalassaemia (iron)
 - Wilson disease (copper)
 - Malignancy

Parathyroid hormone resistance

- Pseudohypoparathyroidism (types Ia, Ib, Ic, II)

Abnormalities of vitamin D metabolism

- Vitamin D deficiency
- Deficient 1α -hydroxylation
 - Renal impairment
 - Acidosis
 - Vitamin D-dependent rickets, type I

Vitamin D resistance

- Vitamin D-dependent rickets, type II
- Severe gluten enteropathy

Other causes

- Acute pancreatitis
- Hyperphosphataemia
- Acute rhabdomyolysis
- Multiple transfusions of citrated blood
- Severe acute illness (e.g. toxic shock syndrome, Gram-negative sepsis)
- Osteoblastic metastases (e.g. prostate, breast)
- Neonatal hypocalcaemia
- Abrupt inhibition of bone resorption (hungry bone syndrome)

sient, with plasma calcium returning to normal in the weeks after surgery. Overt hypocalcaemia can be precipitated in previously asymptomatic patients with partial hypoparathyroidism by therapy with exogenous oestrogens or proton pump inhibitors.

Spontaneous hypoparathyroidism is rare. It may occur as an isolated disorder or be part of a congenital disease complex. The most frequently encountered forms are listed in [Table 6.5](#).

The complex forms are usually the result of mutation in or deletions of genes encoding transcription factors regulating pharyngeal pouch development.

Idiopathic isolated hypoparathyroidism is also likely to have a genetic basis. Mutations in or near the genes *PTH* and *GCMB* (encoding 'glial cell missing'; a master regulator of parathyroid development) have been described in kindreds with both autosomal recessive and autosomal dominant inheritance. A rare X-linked recessive form has also been described.

It is important to distinguish autosomal dominant hypocalcaemia with hypercalciuria (ADHH) from other forms of hypoparathyroidism. This is usually caused by activating mutations in the gene encoding the CaSR and is the genetic and phenotypic converse of familial hypocalciuric hypercalcaemia: PTH concentrations are low in the face of hypocalcaemia, and there is relative hypercalciuria. Most patients with this syndrome are asymptomatic or minimally symptomatic, and do not require treatment. Treatment to raise plasma calcium should not be given to individuals known to have ADHH unless they are symptomatic, because it causes hypercalciuria, which can lead to nephrolithiasis and renal impairment.

Autoimmune hypoparathyroidism may occur alone or in association with additional features, including mucocutaneous candidiasis and adrenal insufficiency, as a component of the autoimmune polyglandular syndrome type 1 (APS-1) or autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED) syndrome. This may be sporadic or familial with an autosomal recessive inheritance pattern. Less commonly, it may also be seen as part of the autoimmune polyglandular syndrome type 2 (APS-2), which is characterized by adult-onset adrenal insufficiency associated with type 1 diabetes and thyroid disease and is believed to be a polygenic disorder with apparent dominant inheritance. The parathyroid glands are an infrequent target for autoimmunity; antibodies directed against the parathyroid CaSR may have a direct pathogenetic role.

Hypoparathyroidism is characterized by hypocalcaemia, with plasma total calcium concentrations as low as 1.25 mmol/L being found. Urine calcium excretion is subnormal and the TmP/GFR (see [p. 111](#)) is increased, leading to hyperphosphataemia. Bone turnover is reduced, as are intestinal calcium absorption and circulating concentrations of both $1,25(\text{OH})_2\text{D}$ and intact PTH. Some PTH immunoassays are not able to distinguish low concentrations from those within the reference range.

Pseudohypoparathyroidism. This term refers to a heterogeneous group of rare conditions characterized by the combination of resistance to the actions of PTH and, variably, other glycoprotein hormones, and Albright hereditary osteodystrophy (short stature, obesity, short

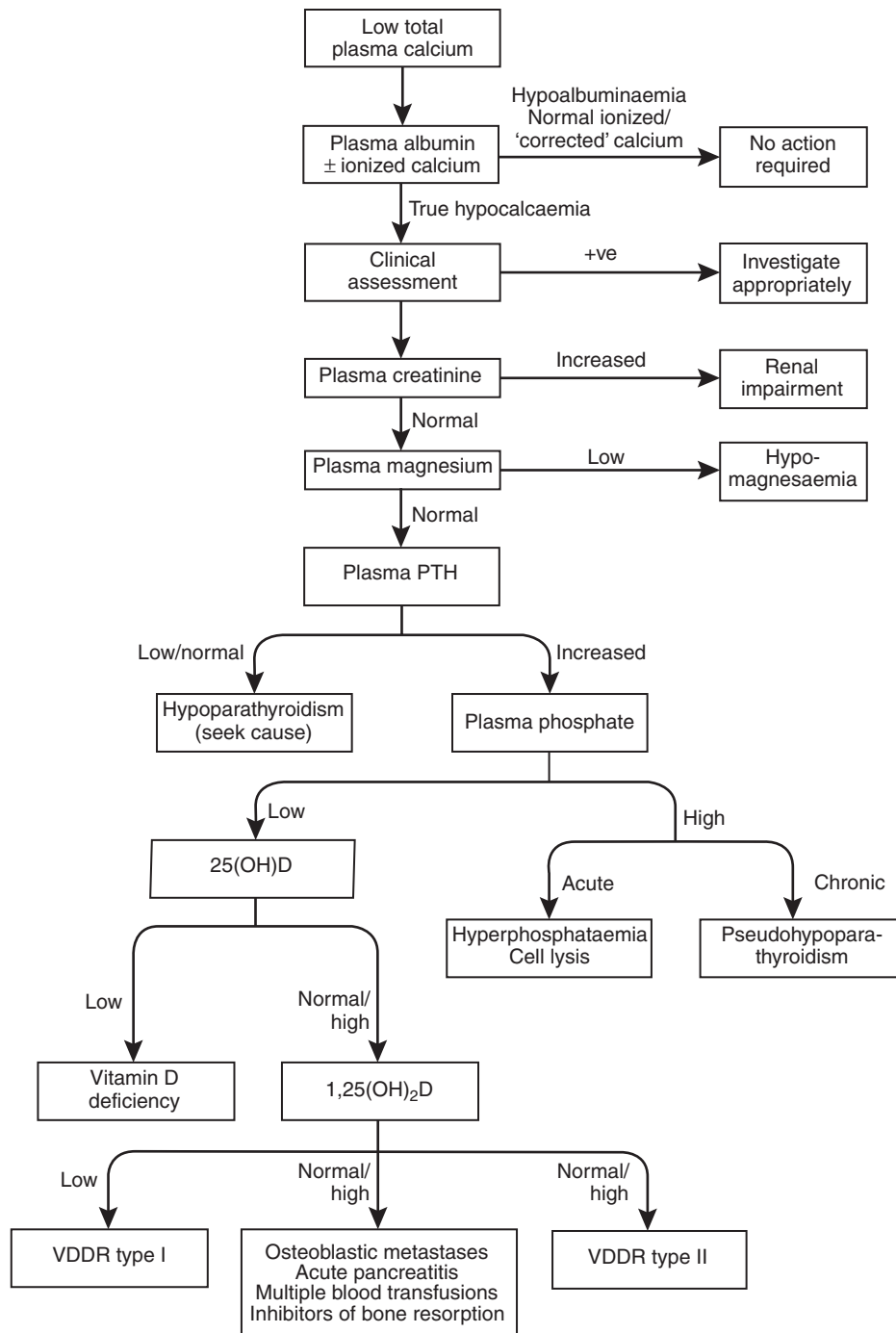


FIGURE 6.6 ■ A flow diagram for investigation of hypocalcaemia. (For abbreviations, see text.)

and the infusion of large quantities of citrate as a result of blood transfusions has a similar effect. Some radiographic contrast dyes contain either citrate or the calcium chelator EDTA, and can also produce hypocalcaemia. These mechanisms, together with magnesium deficiency, which impairs both PTH secretion and action, contribute to the hypocalcaemia observed in 25–50% of patients admitted to intensive care units.

In patients with widespread osteoblastic metastases, modest hypocalcaemia may occur secondarily to increased incorporation of calcium into bone around the secondary deposits. Severe, prolonged hypocalcaemia

may occur in patients following parathyroidectomy for longstanding hyperparathyroidism (hungry bone syndrome). Hypocalcaemia is sometimes seen in the neonatal period, being more common in premature infants and in the offspring of diabetic and hyperparathyroid mothers. Its aetiology is varied (see Chapter 25).

Investigation of hypocalcaemia

Figure 6.6 sets out an approach to the differential diagnosis of hypocalcaemia. Clearly, the first step is to confirm that hypocalcaemia is not just a reflection of low plasma

albumin concentration. An ionized calcium measurement may be necessary, particularly in acutely ill patients in whom calcium binding to albumin or other components of plasma may be abnormal. A number of the causes of hypocalcaemia shown in [Box 6.4](#) are easily distinguished clinically. The acutely ill patient is clearly distinguishable from someone with chronic hypocalcaemia. The family history may be contributory, there may be an abnormal phenotype, evidence of other hormonal abnormalities associated with autoimmune hypoparathyroidism, a past history of thyroid surgery, clinical evidence of osteomalacia etc. Data gained at this point may allow some of the investigations outlined below to be bypassed, or may direct the clinician towards specific diagnoses.

In the absence of such clues, serum creatinine and magnesium should be measured at an early stage, since renal failure and magnesium deficiency are relatively frequent causes of hypocalcaemia. Parathyroid hormone should also be assayed to determine whether parathyroid gland hypofunction is the cause. Serum phosphate measurement is useful, the phosphate concentration being high in hypoparathyroidism, syndromes of PTH resistance and when hyperphosphataemia is the primary disturbance, and (provided renal function is adequate) low in states of abnormal vitamin D activity and/or secondary hyperparathyroidism.

The plan of investigations outlined in [Figure 6.6](#) is only a guide, and some patients will have multiple pathologies, which complicate the interpretation of their biochemistry. This is particularly true in those with acute illness or malignancy.

Some investigations not shown in this scheme may also be of value. A measurement of TmP/GFR is useful in confirming the diagnosis of hypoparathyroidism, and the modified Ellsworth–Howard test in the diagnosis of pseudohypoparathyroidism. Plasma alkaline phosphatase activity is normal or low in hypoparathyroidism, but elevated in most patients with secondary hyperparathyroidism.

Treatment of hypocalcaemia

A detailed discussion of treatment is beyond the scope of this text, but the principles involved will be outlined. Clearly, the underlying condition needs to be sought and, in some cases, appropriate treatment of that will correct the hypocalcaemia. In acute symptomatic hypocalcaemia, intravenous calcium treatment may be necessary. A typical regimen in an adult would be to give 10–20 mL 10% calcium gluconate over 5 min (2.26–4.52 mmol calcium), followed by a continuous intravenous infusion supplying 9–18 mmol calcium in 2 L of fluid over 24 h.

In vitamin D deficiency, treatment should be with vitamin D itself, for reasons of both cost and safety. Repletion of vitamin D does not bypass the body's homeostatic control of plasma calcium concentration, as will occur if 1α -hydroxylated vitamin D metabolites are used. In an adult, adequate vitamin D stores can be achieved by administering 50 000 IU cholecalciferol monthly.

1α -Hydroxylated vitamin D derivatives are preferred in those conditions in which 1α -hydroxylation is impaired and pharmacological doses of vitamin D would otherwise

be needed, such as hypoparathyroidism, pseudohypoparathyroidism and vitamin D-dependent rickets (VDDR) type I. Typical doses for an adult would be 1–4 μ g/day of alfacalcidol or 0.75–2.25 μ g/day of calcitriol. Calcium supplements are not usually required. The aim is to get the plasma calcium concentration just high enough to relieve symptoms, and this is usually achieved around the lower limit of the reference range. Running the plasma calcium concentration higher than this increases the risk of vitamin D toxicity.

PHOSPHORUS METABOLISM

Distribution of body phosphorus

Phosphorus accounts for about 1% of the weight of the elementary composition of the human body; about 23 mol in a 70 kg adult human. Of this, the bulk (85%) is in the skeleton and teeth (largely in the form of hydroxyapatite) and some 14% is located within the cells of the soft tissues. Within different cell types, the phosphorus content may vary from 300 to 1300 mmol/kg. Only 1% of total body phosphorus is present in extracellular fluids ([Box 6.5](#)).

In the blood, phosphorus is present in both organic forms (phosphoproteins, phospholipids etc.) and the inorganic form (as phosphate). The inorganic moiety, which is what is measured routinely in laboratories, exists predominantly in the form of the ions HPO_4^{2-} and H_2PO_4^- in a ratio of about 4:1. Some 15–20% of inorganic phosphate is non-covalently bound to plasma proteins and is not ultrafilterable.

Intracellular phosphorus

Phosphorus is important in the metabolism of all cells. It is a key component of a wide array of biological molecules, including the phospholipids of cell membranes and intracellular organelles, nucleic acids, enzyme cofactors and glycolytic intermediaries. Phosphorylation or dephosphorylation induces dramatic shifts in the activity of many enzymes and cellular signalling proteins. Organophosphate compounds such as adenosine triphosphate (ATP), creatine phosphate and diphosphoglycerate store chemical energy in their high energy phosphate bonds. Hydrolysis of ATP provides the main energy source for many metabolic processes and for

BOX 6.5 Distribution of body and blood phosphorus

Body phosphorus

- Bone ~85%
- Soft tissue cells ~14%
- Extracellular fluid ~1%

Blood phosphorus

- Organic (covalently bound) ~70%
- Inorganic
 - Free (ions, salts) ~25%^a
 - Protein-bound (non-covalent) ~5%^a

^aThe phosphorus content in plasma that is commonly measured and described as phosphate concentration.

muscle contraction. In muscle, ATP is replenished by the donation of phosphate groups from creatine phosphate. In the mitochondria, phosphate-containing proteins play essential roles in the electron transport system. Other phosphate-containing intracellular molecules include cyclic adenosine monophosphate and inositol triphosphate, which act as second messengers.

Although organically-bound phosphorous concentrations within cells are much greater, non-organically bound intracellular phosphate also has important direct effects on cellular energy metabolism. For example, glucose uptake, lactate production and quantity of ATP all vary directly with the intracellular phosphate concentration. The amount of intracellular phosphate is also an important regulator of enzyme activity in the glycolytic pathway. Within red cells, the concentration of 2,3-diphosphoglycerate (2,3-DPG) plays a crucial role in oxygen availability to the tissues. In severe phosphate deficiency, 2,3-DPG synthesis is decreased, which increases the affinity of oxygen to haemoglobin and thus decreases the release of oxygen to the tissues.

Cytoplasmic inorganic phosphate concentrations are similar to those in the extracellular fluid (ECF) (1–1.3 mmol/L), but are not in simple equilibrium. The electrical potential of cells in relation to the ECF means there is a negative potential across the membrane, tending to repel phosphate ions from the cytoplasm. An active process pumps phosphate ions into cells against this membrane potential in order to maintain cytoplasmic phosphate concentrations.

Cytoplasmic inorganic phosphate concentrations change in response to shifts in the balance between the organic phosphorous pool and cytosolic phosphate. Examples of this are in skeletal muscle, when an increase in workload leads to the net degradation of creatine phosphate and generation of cytoplasmic phosphate. In cells in which glycolysis is taking place, such as erythrocytes, a rise in cytoplasmic hydrogen ion concentration stimulates glycolysis. Glycolytic intermediaries, which are organic phosphates, accumulate in the cytoplasm and cause a marked depletion of cytoplasmic phosphate.

Although cells can regulate their steady-state concentrations of phosphate metabolites in the face of changes in extracellular phosphate, severe hypophosphataemia can deplete intracellular phosphate. Large fluxes of phosphate across the cell membrane can, in turn, affect extracellular phosphate concentrations. Various intracellular metabolic disturbances, hormones and hydrogen ion shifts can all cause clinically significant redistribution of phosphate. Examples of mechanisms that permit shifts of phosphate into cells include systemic alkalosis, the stimulation of adrenoreceptors by catecholamines or sympathomimetic agents, and the insulin-mediated entry of glucose into cells.

Phosphate homeostasis

The main sources of phosphate transfer to and from the plasma pool are the intestine, bone, soft tissues and the kidneys. Under steady-state conditions, the net intestinal phosphorus absorption equals the net urine excretion (Fig. 6.7). Small amounts of phosphate (~1 mmol/24 h)

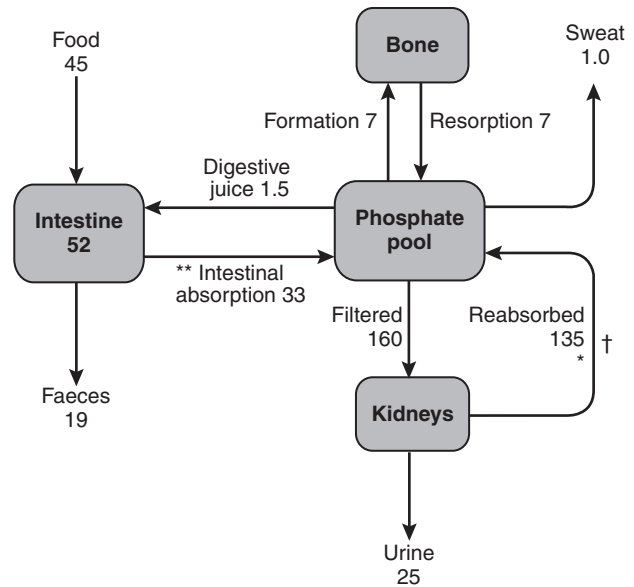


FIGURE 6.7 ■ Representative fluxes (mmol/24 h) in a healthy adult (70 kg body weight) for phosphate. The exchanges with soft tissues are not illustrated in this diagram. For phosphate metabolism the major sites of action of parathyroid hormone (*), FGF23 (†) and 1,25(OH)₂D (**) are indicated. Parathyroid hormone and FGF23 reduce renal tubular reabsorption of phosphate and increase urine phosphate excretion. (Modified from Wilkinson R. Absorption of calcium, phosphorus and magnesium. In: Nordin B E C (ed). Calcium, phosphate and magnesium metabolism. Edinburgh: Churchill Livingstone; 1976, with permission).

may be lost in the sweat. The phosphate content of breast milk is ~1.8 mmol/L, and a fully breast-feeding woman will lose ~1.4 mmol/24 h in the milk.

There are two interrelated mechanisms that react to variations in phosphate status in order to maintain phosphate homeostasis, despite fluctuations either in supply (diet, intestinal absorption) or demand (growth, mineralization, cellular metabolism). These two mechanisms are mediated by:

- 1,25(OH)₂D, synthesis of which occurs in response to hypophosphataemia through the 1 α -hydroxylation of 25(OH)D by the cells of the proximal tubules of the kidney
- fibroblast growth factor 23 (FGF23), synthesis and release of which is stimulated by hyperphosphataemia.

1,25-Dihydroxyvitamin D regulates plasma phosphate primarily by its stimulatory action on intestinal phosphate absorption, whereas FGF23 primarily regulates plasma phosphate by decreasing renal tubular phosphate reabsorption. The two hormonal systems interact: FGF23 production is activated by 1,25(OH)₂D, and in turn potently inhibits the activity of 1 α -hydroxylase required for synthesis of 1,25(OH)₂D. Fibroblast growth factor 23 is most highly expressed in bone, particularly by differentiated osteoblasts and osteocytes, but is also expressed in thymus, brain and heart. It signals through a co-receptor complex comprising the FGF receptor 1 and the β -glycosidase-like protein, klotho, and is catabolized intracellularly by subtilisin-like enzymes that cleave the native protein at a specific site to generate fragments that are thought to be biologically inert. Murine transgenic experiments demonstrate a critical role for FGF23

in phosphate homeostasis: deletion of the *FGF23* gene is accompanied by hyperphosphataemia, enhanced renal tubular reabsorption of phosphate, and elevated concentrations of $1,25(\text{OH})_2\text{D}$; overexpression of the protein causes renal phosphate wasting, decreased concentrations of $1,25(\text{OH})_2\text{D}$ and osteomalacia. Fibroblast growth factor 23 induces phosphaturia by decreasing expression of the two active phosphate transporters in the luminal membrane of the proximal renal tubule, the types 2a and 2c sodium-phosphate co-transporters (NaPi-2a and NaPi-2c, respectively). It further lowers plasma phosphate by inhibiting the activity of renal 1α -hydroxylase and increasing the activity of the $25(\text{OH})\text{D}$ -24-hydroxylase, thereby decreasing synthesis of $1,25(\text{OH})_2\text{D}$, and intestinal phosphate absorption. The concentration of FGF23 rises in response to an increase in plasma phosphate as glomerular function declines, and, by directly promoting an increase in urinary phosphate excretion and indirectly reducing intestinal phosphate absorption, allows the maintenance of normal plasma phosphate concentration in early renal failure.

Fibroblast growth factor 23 is the best characterized of a group of circulating compounds that induce phosphaturia, collectively termed 'phosphatonins'. Other putative phosphatonins, such as the skeletal product matrix extracellular phosphoglycoprotein (MEPE), and secreted frizzled-related protein 4, may be involved in the pathogenesis of pathological states of renal phosphate wasting, hypophosphataemia and osteomalacia (including X-linked hypophosphataemic rickets and oncogenic osteomalacia), but no evidence currently exists for a role of either in physiological phosphate homeostasis.

Dietary phosphate and intestinal absorption

Phosphate is present in a wide range of foodstuffs, so that if the diet is adequate in other nutrients, it is usually also adequate in phosphate. A typical Western diet contains around 0.65 mmol/kg body weight/24 h. There is a considerable capacity to adapt to low phosphate intake, so it is difficult to establish a minimum dietary requirement. Hypophosphataemia resulting solely from inadequate dietary intake is extremely rare. A measurable increment in $1,25(\text{OH})_2\text{D}$ production, one of the adaptive responses, occurs when dietary phosphate is reduced to <0.32 mmol/kg per 24 h.

Phosphate absorption takes place most efficiently in the duodenum and jejunum, though the ileum contributes most in absolute quantity because of its greater length. Approximately 67% of dietary phosphate is absorbed from the jejunum and ileum even under conditions of phosphate repletion, by a passive non-saturable mechanism. Thus, net absorption increases in parallel with dietary phosphate content. The active, $1,25(\text{OH})_2\text{D}$ -dependent, saturable component of phosphate absorption mainly increases jejunal uptake. The bioavailability of intestinal phosphate is reduced by a high calcium intake, as calcium forms insoluble complexes with phosphate in the intestinal lumen. Aluminium salts react similarly. Phosphate is secreted in the digestive juices at a rate of ~0.1 mmol/kg per 24 h, about 67% of which is reabsorbed

from the intestine. Phosphate fluxes are summarized in Figure 6.7.

The renal tubular reabsorption of phosphate

A high proportion of the phosphate in plasma is ultrafilterable and about 75% of filtered phosphate is reabsorbed by the proximal tubules. A further 5–20% is reabsorbed in the distal tubules and/or cortical collecting ducts and the remainder (20–30 mmol/24 h in an adult) appears in the urine.

If phosphate is infused intravenously and the urinary excretion rate of phosphate measured, then above a certain plasma concentration of phosphate all additional increments in the filtered load will be paralleled by the same increment in urinary phosphate; that is, the relationship between the phosphate excretion rate and the plasma phosphate is linear, with a slope that equates to glomerular filtration rate at high (saturating) filtered loads. Extrapolating this line back to where the urine phosphate excretion rate is zero defines the threshold of glomerular filtrate (and hence plasma) phosphate concentration (T_{mP} , tubular maximum for phosphate) that the renal tubular homeostatic mechanism is seeking to maintain, in relation to glomerular filtration rate (GFR). In vivo, this can be determined by calculating the clearance of phosphate relative to that of creatinine on a fasting urine sample and using the nomogram of Walton and Bijvoet (1975), which relates this value to the prevailing plasma phosphate concentration (T_{mP}/GFR). This is performed fasting to avoid changes in plasma phosphate due to food intake, and requires a urine sample and a plasma sample (see Appendix 6.4, for method).

Factors regulating T_{mP}/GFR . A number of endocrine and non-endocrine factors are involved in the regulation of T_{mP}/GFR and hence plasma phosphate (Table 6.7). The most important are FGF23 and PTH. Fibroblast growth factor 23 production is affected by perturbations in plasma phosphate, such that a rise in its concentration stimulates an increase in FGF23 production, which in turn increases urinary phosphate loss. It is not clear whether plasma phosphate concentrations directly affect FGF23 production, or do so by an as yet unknown intermediary pathway. Changes in plasma phosphate do not directly affect PTH secretion, but PTH acts on the renal tubules, to reduce T_{mP}/GFR . Plasma phosphate concentrations are thus usually low when there is an elevation in PTH (e.g. in primary hyperparathyroidism or secondary hyperparathyroidism due to vitamin D deficiency) and renal function is normal. T_{mP}/GFR is increased when PTH concentrations are low, or there is insensitivity to the hormone (e.g. in hypoparathyroidism or pseudohypoparathyroidism). Parathyroid hormone-related peptide has similar effects to PTH on renal phosphate handling. The phosphaturic effects of PTH and PTHrP result from action at multiple sites along the nephrons and are mediated through cAMP production.

The high T_{mP}/GFR seen in childhood and in acromegaly is mediated through insulin-like growth factor

TABLE 6.7 Factors and disorders affecting TmP/GFR

Factor	Increased TmP/GFR	Decreased TmP/GFR
Endocrine	Hypoparathyroidism Pseudohypoparathyroidism Growth hormone/IGF-1 Recovery from vitamin D deficiency Lactation	Hyperparathyroidism PTH-related peptide Corticosteroids Oestrogen replacement therapy
Drugs	Bisphosphonates (particularly etidronate)	Calcitonin Tyrosine kinase inhibitors (imatinib, nilotinib) Diuretics (particularly acetazolamide) Intravenous iron replacement therapy
Natriuresis		Volume loading Single kidney, diuretics Poorly controlled diabetes
Metabolic	Acidosis	Alkalosis
Genetic	Tumoral calcinosis (autosomal recessive) <i>GALNT3</i> (601756) <i>FGF23</i> (605380) <i>KL</i> (604824)	Hypophosphataemic conditions with hypercalciuria and/or osteomalacia/rickets X-linked hypophosphataemic rickets (307800) AD hypophosphataemia (193100) AR hypophosphataemia type 1 (241520) AR hypophosphataemia type 2 (613312) Osteoglophonic dysplasia (166250) McCune–Albright syndrome/fibrous dysplasia (174800) Epidermal nevus syndrome (162900) Hypophosphataemic rickets with hyperparathyroidism (612089)
Acquired		Oncogenic osteomalacia Distal (type 1) renal tubular acidosis Renal Fanconi syndrome Neurofibromatosis

Numbers in brackets refer to Online Mendelian Inheritance in Man (OMIM) catalogue number.

1 (IGF-1). TmP/GFR is increased in thyrotoxicosis and decreased by corticosteroids.

The increase in plasma phosphate seen after the menopause is abolished by oestrogen replacement therapy. There is little evidence that $1,25(\text{OH})_2\text{D}$ greatly affects renal phosphate handling. Calcitonin, given in pharmacological doses, acutely decreases the TmP/GFR. In general, acidosis (whether respiratory or metabolic) reduces renal tubular phosphate reabsorption, whereas alkalosis has the opposite effect. Calcium itself can affect phosphate transport in the kidneys. In healthy subjects, this is difficult to demonstrate because of concurrent changes in PTH, but calcium infusion in hypoparathyroid subjects ultimately increases urinary phosphate losses and reduces plasma phosphate by reducing TmP/GFR.

Various drugs may induce changes in TmP/GFR. In general, these changes are small, the exceptions being glucocorticoids and acetazolamide, which reduce TmP/GFR, and the bisphosphonate, etidronate which, when given at doses of 10 mg/kg, causes a dose-dependent increase in TmP/GFR and plasma phosphate.

Renal phosphate handling may also be affected by haemodynamic changes. Patients with a single kidney (e.g. transplant recipients or live kidney donors) have a high filtered load per nephron and lower TmP/GFR as an adaptation. Phosphaturia also accompanies the natriuresis that follows volume loading, poorly controlled diabetes or diuretic use, because renal phosphate transport is related to sodium transport. Parathyroid hormone is needed for the full expression of this phenomenon.

Disorders of renal phosphate metabolism. There are a number of conditions that share as a central feature a combination of low TmP/GFR, hyperphosphaturia, hypophosphataemia and osteomalacia: the defect may be inherited or acquired. The clinical manifestations depend on the severity of the renal phosphate leak, the degree of response of the 1α -hydroxylase enzyme to the ensuing hypophosphataemia, the age at onset and associated defects. In some instances (e.g. in some patients with idiopathic hypercalciuria), the primary defect appears to be a low TmP/GFR with modest hypophosphataemia and, in response to this, there is an appropriate increase in $1,25(\text{OH})_2\text{D}$ production. Intestinal absorption of calcium and phosphate is enhanced and the extra calcium that is absorbed appears in the urine, predisposing to renal stones.

There are several syndromes of hyperphosphaturia and hypophosphataemia, accompanied by inappropriately low concentrations of $1,25(\text{OH})_2\text{D}$ and osteomalacia, including autosomal dominant hypophosphataemic rickets (ADHR), X-linked hypophosphataemic rickets (XLH) and oncogenic osteomalacia (OOM) (Table 6.7). In most instances, an abnormality in FGF23 metabolism underlies the phenotype.

Oncogenic osteomalacia is a paraneoplastic syndrome, in which tumours of mesenchymal origin elaborate and secrete excess amounts of FGF23, and possibly other phosphatonins, such as MEPE or secreted frizzled-related protein 4. Detection and removal of the tumour reduces FGF23 concentrations and corrects the biochemical and skeletal phenotype. In autosomal dominant hypophosphataemic rickets, affected individuals harbour a mutation in

the gene encoding FGF23, which renders the full-length protein resistant to normal (inactivating) proteolysis. Patients with X-linked hypophosphataemia carry a mutation in the *PHEX* (phosphate-regulating gene with homologies to endopeptidases on the X chromosome) gene, which encodes an endopeptidase that is believed to sequester and stabilize a secreted FGF23-stimulating bone matrix protein. Inactivation of *PHEX* allows accumulation of the matrix-bound FGF23-stimulating protein and an increase in local, and then systemic, FGF23 production. Whether the *PHEX* endopeptidase also inactivates FGF23 in vivo is currently uncertain, but thought to be unlikely.

In all three of these phosphaturic syndromes, circulating concentrations of FGF23 are typically elevated. Circulating concentrations of FGF23 can also be elevated in patients with polyostotic fibrous dysplasia/McCune-Albright syndrome (caused by somatic activating mutations in *GNAS1*) in 50% of whom renal phosphate wasting occurs. Two autosomal recessive conditions are associated with FGF23-dependent hypophosphataemia. In type 1, autosomal recessive hypophosphataemic rickets (ARHR), inactivating mutations of *DMP1*, which encodes an osteocyte protein that restrains FGF23 production, produce a phenotype similar to that of ADHR. Inactivating mutations of *ENPP1*, which encodes an enzyme important in regulating hydroxyapatite deposition, cause type 2 autosomal recessive hypophosphataemic rickets. This is a rare syndrome of generalized arterial calcification of infancy and, in those who survive to adulthood, FGF23-dependent hypophosphataemia.

The recessively inherited condition, hereditary hypophosphataemic rickets with hypercalciuria, is caused by inactivating mutations in the gene *SLC34A3*, which encodes the NaPi-2c transporter. In this hypophosphataemic disorder, which is not mediated by FGF23, chronic hypophosphataemia is accompanied by an appropriate increment in 1α -hydroxylase activity and plasma $1,25(\text{OH})_2\text{D}$. In a few patients, a milder phenotype of hypophosphataemia with nephrolithiasis and osteopenia has been associated with mutations in NaPi-2a. The bone diseases associated with hypophosphataemia are discussed in Chapter 31.

Plasma phosphate concentrations

There are marked age-related changes in plasma phosphate, with the highest concentrations seen in early infancy when growth velocity is highest. Throughout childhood and adolescence, plasma phosphate concentrations remain higher than in adults. In adulthood, a decline in fasting plasma phosphate concentrations is seen in men after the age of 40. In women, there is little change, but a small increment (~ 0.07 mmol/L) occurs after the menopause. Plasma phosphate does not change during pregnancy, but is increased during lactation. The age- and lactation-related changes in plasma phosphate are mediated through changes in TmP/GFR (Table 6.8).

There is marked diurnal variation in plasma phosphate. During the day, concentrations are substantially higher in the mid-afternoon than in the early morning and there is a second peak in the early hours of the morning (see Fig. 6.4). This early morning peak is attenuated with age.

TABLE 6.8 Plasma phosphate concentrations according to age and sex

Age	Sex	Reference range for plasma phosphate (mmol/L)
Last trimester in utero	M, F	0.8–1.4
1 day	M, F	1.0–2.6
1 day–1 month	M, F	1.8–3.2
1–12 months	M, F	1.4–2.1
1–14 years	M, F	1.2–1.7
14–18 years	M, F ^a	1.0–1.6
18–45 years	M, F	0.8–1.4
>45 years	F	0.8–1.4
>45 years	M	0.7–1.2
Lactation	F	1.1–1.5

^aThe decline in plasma phosphate concentrations from adolescent to adult values occurs, on average, two years earlier in girls than boys. Adult concentrations are reached around age 15 in girls and age 17 in boys. This parallels the earlier start and finish to the adolescent growth spurt in girls.

Plasma phosphate rises after food and subsequently falls as phosphate either enters cells or is excreted. The fall in plasma phosphate after a meal is partly attributable to insulin-dependent stimulation of glycolysis, which increases intracellular phosphate utilization.

Hyperphosphataemia

The major causes of hyperphosphataemia are listed in Box 6.6. Hyperphosphataemia may arise through an increase of phosphate input into the blood or decreased

BOX 6.6 Causes of hyperphosphataemia

Pseudohyperphosphataemia

- Haemolysed specimen
- Myeloma
- Delayed separation from red cells

Increased phosphate input

- Intravenous phosphate
- Rectal phosphate

Shift from intracellular to extracellular fluid

- Cell death
 - Tumour lysis syndrome
 - Rhabdomyolysis
 - Malignant hyperpyrexia
 - Heat stroke

Diminished phosphate excretion

- Reduced GFR
 - Acute kidney injury
 - Chronic kidney disease
- Increased TmP/GFR
 - Physiological (normal childhood, lactation, recovery from vitamin D deficiency)
 - Pathological (reduced PTH or PTH resistance), vitamin D toxicity, thyrotoxicosis, acromegaly, etidronate, tumoral calcinosis (genetic conditions, see text)

excretion. The latter may result from a reduction in GFR or an increase in TmP/GFR. Conditions in which the TmP/GFR is raised, generally cause modest hyperphosphataemia (plasma phosphate 1.4–2.4 mmol/L). Tumoral calcinosis in non-uraemic subjects is a rare recessively inherited disorder associated with a high TmP/GFR, elevated 1,25(OH)₂D₃ concentrations and hyperphosphataemia; a biochemical phenotype that is the mirror image of autosomal dominant hypophosphataemic rickets. It is most commonly caused by inactivating mutations in the *GALNT3* gene. *GALNT3* encodes a glycosyltransferase known as ppGalNac-T3. This enzyme O-glycosylates the subtilisin-like proprotein convertase recognition site in the FGF23 protein, thereby protecting FGF23 from proteolytic cleavage and permitting intact FGF23 to be secreted. In the absence of this enzyme, most FGF23 is processed into inactive fragments before secretion, leading to low or undetectable circulating concentrations of biologically active FGF23. Tumoral calcinosis can also result from mutations in the genes encoding FGF23 or *klotho*, leading to impaired secretion of FGF23, and resistance to the actions of FGF23, respectively. A feature of tumoral calcinosis is the presence of large periarticular calcific masses, which are subject to breakdown and chronic inflammation.

Renal impairment, whether acute or chronic, can cause hyperphosphataemia, but this does not usually occur until the GFR falls below 30 mL/min. As renal function deteriorates further, plasma phosphate rises sharply and may reach 6 mmol/L in established renal failure. Other abnormalities of phosphate metabolism that present early in the course of chronic kidney disease are important in the aetiology of hyperparathyroidism and are discussed further in Chapter 31.

The diet rarely contains sufficient phosphate to cause significant postprandial hyperphosphataemia, but excessive input of phosphate can occur with the use of intravenous phosphate salts or phosphate enemas. The latter is particularly liable to occur with children unless the dose of phosphate in the enema is reduced. Massive cell death with release of intracellular phosphate can cause hyperphosphataemia. Examples of the latter include rhabdomyolysis and the tumour lysis syndrome, seen, for example, after the initiation of chemotherapy in childhood leukaemia. It should be emphasized that, as with hypercalcaemia, an increased input of phosphate into the circulation is much more likely to cause severe hyperphosphataemia if there is coexistent renal impairment and thus a reduced capacity to excrete phosphate. This is often the case in critically ill patients.

Plasma phosphate may be artefactually raised (by ~30%) in very haemolysed samples or if separation from red cells is delayed. Pseudohyperphosphataemia has also been described with certain chromogenic assay methods of automated analysis in patients with paraproteinaemia or hypertriglyceridaemia.

Consequences of hyperphosphataemia

No specific symptoms are directly attributable to hyperphosphataemia. However, when the product of the calcium and phosphate concentrations in blood exceeds the

solubility product (approximately 4.85×10^{-6} molar units), soft tissue (metastatic) calcification occurs. This is seen most frequently in chronic kidney disease, where calcification occurs in blood vessels, the skin, heart, lungs, kidneys, conjunctivae and around joints. Interstitial calcification of the kidneys is of particular importance, since it impairs renal function further and reduces the ability to excrete phosphate loads. The deposition of calcium phosphate salts in the skin is thought to contribute to the pruritus of chronic uraemia, a symptom that may improve with better control of hyperphosphataemia. The syndrome of progressive ischaemic skin ulceration (calciophylaxis) in patients with established renal failure has also been attributed to calcium phosphate deposition. In this syndrome, ulcers spread rapidly with secondary infection of the necrotic tissues. Histologically, there is extensive arterial and arteriolar calcification and thrombosis. These lesions do not respond to local therapy, but do heal after parathyroidectomy (which reduces the [calcium]×[phosphate] product acutely). Urgent parathyroidectomy is indicated in this condition.

If plasma phosphate rises acutely, then plasma calcium concentrations fall by a mass action effect; this is seen, for example, with the tumour lysis syndrome and rectal or intravenous phosphate administration. This was the basis of the former use of intravenous phosphate in the management of malignant hypercalcaemia. However, the perils of soft tissue calcification induced by this treatment outweigh any benefit.

Diagnostic approach to hyperphosphataemia

The clinical history should reveal excessive exogenous phosphate supply (intravenous or rectal). Massive cell death in, for example rhabdomyolysis or the tumour lysis syndrome, should also be evident from the clinical setting. Rhabdomyolysis can be confirmed by measurement of serum creatine kinase activity.

Diminished excretion of phosphate due to severe renal impairment is evident from measurement of the plasma creatinine. The most convenient method of detecting changes in TmP/GFR is to use the nomogram of Walton and Bijvoet (see Appendix 6.4, below). Other measures of urinary phosphate clearance have theoretical failings compared with the TmP/GFR. In interpreting the result, it is important to bear in mind the normal age-related changes in this index. There is no diagnostic value in measuring the 24 h urine phosphate excretion which, in steady-state conditions, reflects the product of dietary phosphate content and fractional phosphate absorption.

Therapeutic approach to hyperphosphataemia

Hyperphosphataemia due to a raised TmP/GFR is rarely high enough to require specific action. Phosphate control is, however, important in patients with established renal failure, both to limit damaging metastatic and vascular calcification and to control the development of hyperparathyroidism. There are various salts of aluminium, magnesium and calcium which, when administered orally, bind phosphorus in the intestine and

limit its absorption. Aluminium hydroxide is the most effective of these agents, but significant quantities of aluminium, which has toxic effects on bone, may be absorbed. Calcium-based phosphate binders (such as calcium acetate or calcium carbonate) are generally used in preference to aluminium salts. However, calcium-based phosphate binders tend to worsen hypercalcaemia in dialysis patients, and have been linked to adverse vascular outcomes, so calcium-free phosphate-binding agents, such as lanthanum carbonate and sevelamer hydrochloride are increasingly used. In dialysis patients, inadequate phosphate control may be due to a number of factors (see Box 31.6).

Acute kidney injury can complicate the tumour lysis syndrome and is attributable to phosphate and urate released from cells. Maintaining a brisk saline diuresis at the time chemotherapy is begun reduces the risk of developing acute kidney injury.

Hypophosphataemia

Mechanisms

Hypophosphataemia can arise by any one or combination of three mechanisms:

- inadequate phosphate absorption from the intestine
- shifts of phosphate from extracellular fluid into cells
- abnormal urinary phosphate losses (Table 6.9).

As indicated earlier, it is very rare that the diet is so inadequate in phosphate that, on its own, it can cause hypophosphataemia. However, chronic malnutrition

can result in a deficit of whole body phosphate such that, when conditions occur in which there is a shift of phosphate from the extracellular to the intracellular compartment, such as during refeeding, profound hypophosphataemia may develop. Inadequate dietary phosphate may thus be a predisposing factor to, rather than a direct cause of, hypophosphataemia. Clinically significant hypophosphataemia due to the use of phosphate-binding antacids has been described, but this is now rare because of the availability of alternative therapies for peptic ulcer disease.

Hypophosphataemia can result from the acute shift of phosphate into cells. Less than 1% of total body phosphate is in the extracellular space, so small shifts of no more than 6 or 7 mmol can induce significant changes in plasma phosphate concentrations. If this occurs without prior intracellular phosphate depletion, then the condition is usually benign and self-limiting. The most frequently encountered conditions in which this phenomenon occurs are the following.

- *Refeeding syndrome.* Providing sufficient calories, protein and other nutrients to malnourished patients allows accelerated tissue accretion. The cellular uptake and utilization of phosphate increases, initially for the purpose of phosphorylation of glucose and fructose, and ATP synthesis. When insufficient quantities of phosphate are provided, an acute state of severe hypophosphataemia, with intracellular phosphate depletion, occurs. The clinical and metabolic consequences can be serious. Examples of situations in which this mechanism can operate, include the treatment of diabetic ketoacidosis and during parenteral nutrition.

TABLE 6.9 Mechanisms underlying common causes of hypophosphataemia

Aetiology of hypophosphataemia	↑ Cellular uptake			↓ Gastrointestinal input		↑ Urinary loss	
	Respiratory alkalosis	Refeeding	Muscle	Inadequate intake	Poor absorption	↓ TmP/GFR	
Primary hyperparathyroidism						•	Plasma phosphate typically 0.35–0.80 mmol/L
Vitamin D deficiency with secondary hyperparathyroidism				•	•	•	
Post renal transplant					•	•	
Hypophosphataemic rickets or osteomalacia					(•) ^a	•	
Antacid consumption					•	•	
Sepsis, burns	•			•	•	•	
Alcoholism				•		•	
Sympathomimetics (e.g. β-agonists)			•				
Salicylate poisoning	•						
Early asthmatic crisis	•		(•) ^b				
Alcohol – withdrawal	•	•		•			Plasma phosphate may be <0.35 mmol/L
Diabetic ketoacidosis – recovery		•				(•) ^c	
Hyperalimentation		•		•			
Hepatic failure	•	•		•	•		
Long distance running			•				

^aContribution of phosphate malabsorption depends on 1,25(OH)₂D response to hypophosphataemia.

^bDue to β-agonist use.

^cPhosphaturia associated with glycosuria.

- *Respiratory alkalosis.* Hyperventilation due to any cause will reduce the carbon dioxide content of the blood and cells. Intracellular hydrogen ion concentration falls, which activates the enzyme phosphofructokinase, which in turn accelerates phosphorylation of glucose. Phosphate is taken up rapidly from the plasma and hypophosphataemia results. Examples of clinical situations in which this mechanism may occur include liver failure, severe burns, salicylate poisoning, alcoholic ketoacidosis, alcohol withdrawal and sepsis. Hyperventilation in these situations may be the result of either direct metabolic effects or simply fear and pain. It is commonly observed in hospitalized patients, but is typically mild and transient and does not require specific treatment; correction of the underlying cause of hyperventilation also corrects the hypophosphataemia.
- *Increased muscle uptake.* Prolonged intense exercise can increase uptake of phosphate into muscle in an attempt to replenish creatine phosphate stores. This has caused profound hypophosphataemia in marathon runners.

Increased urinary phosphate loss is usually a chronic phenomenon resulting from a low TmP/GFR and gives rise to moderate hypophosphataemia and osteomalacia (see above, and Chapter 31). In uncontrolled diabetes, osmotic diuresis causes loss of phosphate in the urine, and this contributes to the occurrence of hypophosphataemia once insulin therapy is begun. A reduction in TmP/GFR also contributes to the hypophosphataemia that can occur in paracetamol overdose, even in the absence of significant hepatic damage. A reduction in TmP/GFR, along with other renal tubular function abnormalities, may be seen in patients with severe alcohol dependency. It reverses on abstinence.

Consequences of hypophosphataemia

Mild to moderate hypophosphataemia (0.35–0.80 mmol/L in an adult) is not harmful in the short term, but if chronic it may induce osteomalacia. When plasma phosphate concentrations fall below 0.35 mmol/L, the syndrome of acute phosphate deficiency may develop. The major manifestations are listed in Table 6.10.

Diagnostic approach to hypophosphataemia

The acute phosphate deficiency syndrome often arises in patients who are critically ill from other causes, and the symptoms of phosphate deficiency may mimic features of these illnesses. Examples of this include deteriorating mental function simulating encephalopathy in the hepatic failure patient, and tremor and irritability simulating alcohol withdrawal symptoms. The detection of severe phosphate deficiency thus depends on anticipating the circumstances in which it might occur and requesting laboratory confirmation. In these circumstances, measurement of plasma phosphate is usually sufficient, but assessments of recent dietary and intravenous phosphate intake may also be helpful. In all patients with severe acute hypophosphataemia, phosphate disappears from the urine so little information can be gained from its measurement.

In the case of chronic hypophosphataemia causing osteomalacia, the critical investigation is determination of the TmP/GFR. An inappropriately low TmP/GFR should prompt evaluation for one of the causes of renal phosphate wasting, discussed under Disorders of renal phosphate metabolism (see p. 112). Measurement of serum concentrations of FGF23 can be helpful in identifying FGF23-dependent causes of hypophosphataemia. Development of FGF23 immunoassays is continuing; there is an assay that measures the full-length peptide only and a C-terminal assay that recognizes both full length and the cleaved C-terminal fragment.

Nuclear magnetic resonance has been used to detect changes in intracellular phosphate metabolism and has proved particularly valuable in research into muscle diseases. The technique is not, however, generally available for clinical practice. Isotopic tests of intestinal phosphate absorption have also been used in research, but have little practical clinical application.

Therapeutic approach to hypophosphataemia

Mild to moderate acute hypophosphataemia secondary to redistribution (plasma phosphate 0.35–0.80 mmol/L in adults) is usually transient and requires no specific

TABLE 6.10 The acute phosphate deficiency syndrome

System affected	Mechanism/site	Clinical consequence
Haemopoietic system	↓ 2,3-DPG in red cells, reduced O ₂ release to peripheral tissues	Tissue hypoxia
	↑ Red cell fragility Abnormal leukocyte function Abnormal platelet function	Haemolysis (plasma phosphate <0.1 mmol/L) ↓ Resistance to infection ↓ Platelet survival, abnormal clotting
Striated muscle	↓ Contractility of diaphragm, heart, skeletal muscles	Respiratory failure ↓ Stroke work Stiffness, weakness, debility
	Rhabdomyolysis	Muscle pain, weakness ↑ Plasma creatine kinase
Nervous system	Brain	Lethargy, confusion, irritability, dysarthria, tremor, seizure, coma
	Peripheral nerves	Paraesthesiae
Gastrointestinal	↓ conduction velocity	
	Smooth muscle	Gastric atony, ileus

treatment. Severe hypophosphataemia (<0.35 mmol/L) occurs only when there has been a cumulative net loss of ~100 mmol phosphate. Symptoms of hypophosphataemia appear when net losses reach ~300 mmol. However, the deficit in total body phosphate cannot be predicted reliably from the plasma phosphate concentration.

Treatment may be given orally in the form of sodium or potassium hydrogen phosphate (100 mmol phosphate daily in divided doses for a week), but often with critically ill patients, intravenous replacement is necessary, particularly in those with symptomatic acute hypophosphataemia. Intravenous therapy can safely be undertaken using monobasic potassium phosphate (KH_2PO_4) in a dose of 50–100 $\mu\text{mol/kg}$ body weight in half normal saline by continuous infusion over 12 h. This can be repeated every 12 h with monitoring of plasma calcium, phosphate and potassium. Intravenous administration can be stopped when plasma phosphate has risen above the symptomatic watershed of 0.35 mmol/L, assuming the rest of the deficit can be replaced orally.

Prevention is appropriate for at-risk patients. For example, many patients receiving parenteral nutrition require about 0.5 mmol phosphate/kg body weight/24 h, and occasionally more. Malnourished alcoholic patients receiving intravenous glucose should receive phosphate supplementation too.

The complications of intravenous phosphate administration include hyperphosphataemia, hypocalcaemia, hyperkalaemia and acidosis. Oral phosphate is a laxative that may induce diarrhoea.

MAGNESIUM METABOLISM

Most magnesium is in the mineral phase of bone or within the cells of the soft tissues. After potassium, it is the most abundant intracellular cation. Within cells, magnesium is a cofactor in over 300 enzymatic reactions, involving energy metabolism, control of various calcium and potassium channels, membrane stabilization and neuromuscular excitability, protein and nucleic acid synthesis and oxidative phosphorylation. It is of particular importance in those processes involving the formation and utilization of ATP. All enzymatic reactions involving ATP have an absolute requirement for magnesium. Less than 0.5% of total body magnesium is present in the plasma, although, for practical purposes, measurements of plasma ionized or total magnesium are the only widely used methods for determining magnesium status. However, such measurements may not accurately reflect whole body magnesium status. The body distribution of magnesium in an adult is summarized in Table 6.11. The magnesium content of infants is lower than adults (approximately 10 mmol/kg body weight, compared with 16 mmol/kg body weight in adults).

Plasma magnesium

In plasma, ~60% of the total magnesium is in the ionized form and ~15% complexed (with phosphate, citrate or bicarbonate), the remainder (~25%) being protein-bound,

TABLE 6.11 Magnesium content and distribution in a typical 70 kg adult

	Total (mmol)	Concentration
Bone	600	100 mmol/kg dry weight
Intracellular		
Soft tissues	500	12–18 mmol/kg
Red cells	5.5	2.7 mmol/L
ECF		
Interstitial fluid	7.0	0.7 mmol/L
Plasma	2.5	0.8–1.0 mmol/L
Ionized	1.5	0.5–0.6 mmol/L

mainly to albumin. Plasma magnesium concentrations can be adjusted for albumin concentration by the formula:

$$\text{Mg}_C^{2+} = \text{Mg}_T^{2+} + 0.005(40 - \text{Alb});$$

where Mg_C^{2+} is the corrected value and Mg_T^{2+} is the measured total value. The ionized and complexed magnesium fractions are ultrafilterable. The precise distribution of magnesium in the various plasma fractions may, as in the case of calcium, be altered by changes in protein and hydrogen ion concentrations. Plasma magnesium concentrations do not show any major changes with age or between the sexes, apart from a small increment at the menopause of ~0.06 mmol/L. Plasma magnesium falls to premenopausal concentrations with oestrogen replacement, in the same way that calcium does.

Magnesium homeostasis

Magnesium is widely distributed in foods. Being the mineral ion of chlorophyll, green vegetables are an important source. Daily intakes range from 6 to 20 mmol, with a median of ~12 mmol. Hard water contains an appreciable amount of soluble magnesium (up to 5 mmol/L), which may be more available than some of the intracellular magnesium in particular foods. Food processing can remove much of the magnesium from cereals and carbohydrate foodstuffs. The wide reference range for 24 h urinary magnesium excretion (2.0–7.5 mmol) largely reflects different magnesium intakes.

Of the typical dietary intake of 12 mmol, 6 mmol is absorbed, mainly in the small intestine. There are two separate transport systems. One, an active transport, is saturated at low intraluminal concentrations; the other is a passive diffusion mechanism that absorbs a constant fraction (~7%) of ingested magnesium. Some magnesium can be absorbed from the large bowel, as demonstrated by the occurrence of hypermagnesaemia after the use of magnesium-containing enemas. About 2 mmol magnesium is secreted into the intestine, thus the net daily intestinal absorption is ~4 mmol, which is balanced by excretion of this amount into the urine.

Approximately 75% of the total plasma magnesium is filtered through the glomeruli. Of that, ~20% is reabsorbed in the proximal tubules by paracellular absorption. The majority of the ultrafilterable magnesium (~65%) is reabsorbed in the cortical thick ascending limb of the loop of Henle. This is a passive paracellular process, driven by a positive transepithelial

voltage gradient and mediated by a tight-junction Mg^{2+} pathway, which involves the molecules claudin-16 and claudin-19.

The distal convoluted tubules regulate the final urinary magnesium excretion by active transcellular reabsorption of Mg^{2+} (~5–10% of ultrafilterable magnesium). Active magnesium reabsorption is a multistep process. A negative membrane potential maintained by the voltage-gated K^+ channel $Kv1.1$ provides the driving force for Mg^{2+} transport across the apical, epithelial channel, TRPM6 (transient receptor potential cation channel M6). Epidermal growth factor regulates active Mg^{2+} transport through TRPM6 in a paracrine/autocrine manner. The Na^+,K^+ -ATPase situated in the basolateral membrane generates a local negative membrane potential helping to drive Mg^{2+} across the TRPM6 channel in the basolateral membrane. The transcription factor HNF1 β (hepatocyte nuclear factor 1 β) affects the expression of the regulatory protein γ -subunit that binds and modulates Na^+,K^+ -ATPase. There is a probable active extrusion mechanism, either via a Mg^{2+} pump or a Na^+-Mg^{2+} exchanger, with the transmembrane protein CNNM2 (cyclin M2) a likely candidate.

There is no significant reabsorption in the collecting ducts. Between 3% and 5% of the filtered magnesium finally appears in the urine. The kidney has a maximal limit for tubular reabsorption ($TmMg$), above which all the ultrafilterable magnesium is excreted. Calculating either the $TmMg$ or the fractional urinary magnesium excretion from magnesium infusion experiments can be helpful in characterizing genetic hypomagnesaemic disorders.

Renal magnesium clearance is increased by osmotic diuretics, loop diuretics and thiazides. The kidneys of the premature newborn differ from those of adults in that the proximal tubules can reabsorb up to 70% of the ultrafilterable magnesium.

There are clear homeostatic mechanisms for regulating magnesium status, although the magnesium 'sensor' has not been characterized. The kidneys appear to be the prime organs in the fine regulation of magnesium homeostasis. In times of magnesium deprivation, the kidneys can reduce magnesium excretion to <1 mmol/24h. There is intestinal adaptation too, with the proportion of dietary magnesium absorbed higher (up to 75%) when the diet is magnesium-depleted, but lower (down to 25%) when the load is greater. Intestinal magnesium absorption can be inhibited by binding agents such as cellulose phosphate. Faecal magnesium output is normally <15 mmol/24h.

Magnesium appears to regulate parathyroid hormone secretion in a manner similar to calcium, but is only one-half to one-third as potent. Parathyroid hormone secretion is thus stimulated by modest hypomagnesaemia and suppressed by hypermagnesaemia. Paradoxically, profound hypomagnesaemia actually inhibits parathyroid hormone secretion. Although PTH may enhance the renal tubular reabsorption of magnesium, it is probably of little importance in the overall regulation of magnesium metabolism.

Magnesium losses through sweat may be great. At high temperatures, 10–15% of total magnesium output can be

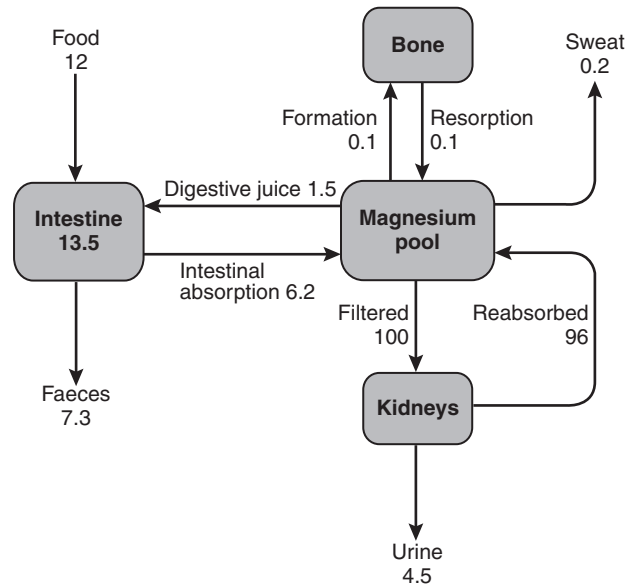


FIGURE 6.8 ■ Representative fluxes (mmol/24h) in a healthy adult (70 kg body weight) for magnesium. The exchanges with soft tissues are not illustrated in this diagram. (Modified from Wilkinson R. Absorption of calcium, phosphorus and magnesium. In: Nordin B E C (ed). Calcium, phosphate and magnesium metabolism. Edinburgh: Churchill Livingstone; 1976, with permission).

found in the sweat. Typical magnesium fluxes in an adult are illustrated in Figure 6.8. The magnesium content of breast milk is ~1.6 mmol/L, and a breast-feeding woman may lose up to 1.2 mmol/24h through this route.

Hypomagnesaemia

Causes

Hypomagnesaemia may arise through inadequate absorption (or intake), by excessive urinary or gastrointestinal losses, or by redistribution of magnesium from extracellular to intracellular sites. Hypomagnesaemia in adults is usually an acquired condition, the major causes of which are listed in Box 6.7. It is important to recognize that several drugs can cause symptomatic hypomagnesaemia, and that withdrawal of the drug concerned will resolve the problem. Most drugs associated with hypomagnesaemia affect renal tubular magnesium reabsorption. The exceptions are proton pump inhibitors (PPI), which probably affect intestinal transport or increase intestinal loss. This is a class effect and hypomagnesaemia will recur if a second PPI is substituted for the first. In adults, severe symptomatic hypomagnesaemia (<0.35 mmol/L) is seen in patients having platinum-based chemotherapy or taking proton pump inhibitors, or who are very unwell in intensive care settings.

There are a number of familial disorders (predominantly affecting renal tubular magnesium handling) that have helped understanding of magnesium homeostasis (Table 6.12). The recessively-inherited forms tend to be more severe and present in infancy or childhood. The dominantly inherited forms tend to be milder (although hypomagnesaemia can be severe with *KCNA1* mutations, plasma calcium is often unaffected).

BOX 6.7 Acquired causes of hypomagnesaemia**Inadequate intake**

- Alcoholism
- Protein calorie malnutrition
- Prolonged infusion or ingestion of low Mg solutions or diet

Malabsorption

- Inflammatory bowel disease
- Proton pump inhibitor treatment (may be excessive gut loss rather than malabsorption)^a
- Gluten enteropathy
- Intestinal bypass
- Radiation enteritis

Renal tubular dysfunction

- Alcoholism
- Hyperaldosteronism
- Hyperparathyroidism
- Post renal obstruction
- Post renal transplant
- Nephrotoxic drugs (amphotericin^a, cisplatin^a, ciclosporin^a, cetuximab, amikacin, gentamicin, laxatives, pentamidine, tobramycin, tacrolimus and carboplatin)
- Potassium depletion
- Diuretics (non-potassium-sparing)
- Osmotic diuresis

Intracellular shift

- Post-myocardial infarction
- Post-parathyroidectomy
- Recovery from diabetic ketoacidosis
- Recovery from starvation (refeeding syndrome)
- Acute pancreatitis

^aMost common causes of severe drug-induced hypomagnesaemia.**Consequences**

Symptoms directly attributable to hypomagnesaemia occur at plasma concentrations <0.5 mmol/L and include anorexia, nausea, tremor, apathy, depression, agitation and confusion. Severely hypomagnesaemic patients usually have coexistent hypokalaemia and a cellular potassium deficit; probably because of reduced activity of the magnesium-dependent Na⁺,K⁺-ATPase system of cell membranes. Potassium is lost from cells into the ECF and is then lost in the urine because magnesium deficiency also impairs the renal conservation of potassium. If severe hypomagnesaemia is prolonged then hypocalcaemia may also develop, principally because of impaired parathyroid hormone secretion. With the development of hypocalcaemia, neuromuscular signs such as muscle weakness, tremor and twitching, a positive Chvostek sign, delirium, convulsions and ultimately coma may occur.

The effects of hypomagnesaemia on the excitability of nerve, muscle and cardiac cells appear not to be direct since, as far as is known, magnesium is not involved in the normal action potential, in conduction or contraction. The effects appear to be mediated by inhibition of the Na⁺,K⁺-ATPase enzyme and, in practice, it is difficult to distinguish direct neuromuscular and cardiac effects of

severe hypomagnesaemia from those secondary to alterations in potassium and calcium metabolism.

Cardiac effects. Hypomagnesaemia of a more modest degree (0.5–0.8 mmol/L) is not uncommon and, while it is not associated with obvious symptoms, it is not necessarily benign. In patients admitted to coronary care units, there is a relationship between hypomagnesaemia and the incidence of serious dysrhythmias (including multifocal atrial tachycardia, ventricular tachycardia, ventricular fibrillation and torsades de pointes). However, hypomagnesaemia is strongly associated with hypokalaemia, and this is itself a potent cause of dysrhythmias. The association between low plasma concentrations of magnesium and potassium is again explicable by reduced membrane ATPase activity, a loss of intracellular potassium and subsequent urinary losses. The ratio of intra- to extracellular potassium, which largely determines the resting membrane electrical potential, is reduced with a consequent increase in electrical excitability. Hypokalaemia can be refractory to replacement therapy while magnesium deficiency persists. Some cardiac dysrhythmias are less responsive to antidysrhythmic therapy in the presence of hypomagnesaemia. In addition to the dysrhythmias, magnesium depletion is associated with other electrocardiographic abnormalities, including PR and QT interval prolongation and T-wave flattening. Following acute myocardial infarction, plasma magnesium concentrations typically fall and are at their lowest 12–24 h after admission to hospital. Drug therapy, particularly non-potassium-sparing diuretics, may exacerbate this hypomagnesaemia by increasing urinary losses.

Diagnostic approach to hypomagnesaemia

Plasma magnesium concentration remains the only widely available measure of magnesium status. Attempts to assess intracellular magnesium more reliably have included the measurement of erythrocyte, monocyte and muscle cell concentrations, but these experimental approaches are not widely available. Hypomagnesaemia is common in the critically ill, occurring in up to 50% of such patients. However, only about 25% have low ionized magnesium concentrations.

A magnesium retention test, which aims to detect a total body deficit, has been proposed. The urinary excretion of magnesium is measured over 24 h following the parenteral administration of 0.1 mmol/kg magnesium (in the form of magnesium sulphate). Normally 75–80% of the dose is excreted, but less if there is magnesium deficiency (see [Appendix 6.5](#), for method).

To assess renal magnesium handling in patients with suspected renal tubular abnormalities, the fractional urinary magnesium (FE_{Mg}) can be estimated on fasting urine (U) and plasma (P) samples:

$$FE_{Mg} (\%) = \frac{U_{Mg} \times P_{creat}}{(0.75 \times P_{Mg}) \times U_{creat}} \times 100$$

Values >2%, while the patient is hypomagnesaemic, suggest renal magnesium wasting. Renal tubular

TABLE 6.12 Genetic causes of hypomagnesaemia

Disorder	OMIM ^a	Inheritance locus	Gene	Mutated protein	Mechanism	Phenotype
Hypomagnesaemia 1 (HOMG1)	602014	AR 9q21	<i>TRPM6</i>	Transient receptor potential channel M6 (TRPM6)	Reduced absorption of Mg from distal convoluted tubule and impaired intestinal Mg absorption	Neonatal convulsions, tetany
Hypomagnesaemia 2 (HOMG2)	154020	AD 11 q23	<i>FXYD2</i>	Na ⁺ ,K ⁺ -ATPase γ subunit	Reduced absorption of Mg from distal convoluted tubule	Often asymptomatic, hypocalciuria
Hypomagnesaemia 3 (HOMG3)	248250	AR 3q27	<i>CLDN16</i>	Paracellin-1 (Claudin-16)	Reduced paracellular absorption of Mg (and Ca) from thick ascending limb	Neonatal convulsions, polyuria, nephrocalcinosis
Hypomagnesaemia 4 (HOMG4)	611718	AR 4q25	<i>EGF</i>	Epidermal growth factor	Impaired activation of TRPM6 and reduced absorption of Mg from distal convoluted tubule	Learning disability, epilepsy
Hypomagnesaemia 5 (with ocular involvement) (HOMG5)	248190	AR 1p34	<i>CLDN19</i>	Claudin-19	Reduced paracellular absorption of Mg (and Ca) from thick ascending limb	Nephrocalcinosis, renal impairment, macular colobomata
Hypomagnesaemia 6 (HOMG6)	613882	AD 10q24	<i>CNNM2</i>	CNNM2	Basolateral renal tubular epithelial transport	
Gitelman syndrome	263800	AR 16q 13	<i>SLC12A3</i>	Thiazide-sensitive Na-Cl co-transporter	Reduced absorption of Mg from distal convoluted tubule	Hypokalaemia, tetany precipitated by non-specific illness, hypocalciuria
Hypomagnesaemia, hypertension and hypercholesterolaemia	500005	Maternal (mitochondrial)	tRNA(Ile)	tRNA isoleucine	Unknown	Migraine, deafness, hypertrophic cardiomyopathy in adults
Hypocalcaemia with hypercalciuria	601199	AD 13q 13	<i>CASR</i>	Calcium-sensing receptor ^b	? Reduced renal tubular reabsorption	Usually asymptomatic
Myokymia with hypomagnesaemia	160120	AD 12p 13	<i>KCNA1</i>	Voltage-gated potassium channel, Shaker-related subfamily 1 (Kv1.1)	Impaired renal tubular reabsorption	Myokymia, episodic ataxia. Plasma calcium usually unaffected despite significant hypomagnesaemia
MODY 5	125853	AD 17q12	<i>HNF1B</i>	Hepatic nuclear factor 1B	Impaired renal tubular reabsorption	Diabetes with renal cysts

^aOnline Mendelian Inheritance in Man phenotype catalogue number.

^bIndicates an activating mutation; all others are thought to be inactivating.

magnesium handling can be assessed more rigorously by magnesium infusion testing. It is possible to combine an infusion test to determine TmMg with a magnesium retention test to determine if there is a magnesium deficit (see [Appendix 6.6](#), for a suggested protocol).

Where hypomagnesaemia is due to inadequate intake or malabsorption, urinary magnesium falls rapidly to <0.7 mmol/24h from the usual 2–7 mmol/24h.

The main diagnostic step must remain thinking of hypomagnesaemia as a potential problem and making a request for plasma measurement. Studies comparing

physician-initiated testing and ‘routine’ measurement have shown that hypomagnesaemia is very common in hospitalized patients, that it is frequently associated with hypokalaemia and that only about 10% of hypomagnesaemic patients are identified by physician-initiated requests. This has been used as an argument for including magnesium among the routine hospital analyses of plasma ‘electrolytes’.

The ingestion of large quantities of magnesium (in the form of antacids or laxatives) can induce diarrhoea. In magnesium-related diarrhoea, the faecal magnesium

BOX 6.8 Conditions that may benefit from magnesium supplementation

- Ventricular tachyarrhythmias
- Hypokalaemia unresponsive to potassium supplementation
- Neurological manifestations of hypomagnesaemia
- Hypomagnesaemia without clinical manifestations
- Diuretic therapy
- Hypocalcaemia unresponsive to calcium supplementation

From Reinhart R A 1988 Magnesium metabolism. Archives of Internal Medicine 148:2415–2420.

content is high (>15 mmol/24h). This measure is of occasional value in the diagnosis of factitious diarrhoea.

Therapeutic approach to hypomagnesaemia

Conditions that may benefit from magnesium repletion are listed in Box 6.8. The amount and rate of magnesium administration depends on the cause and severity of the magnesium depletion and on renal and intestinal function.

Symptomatic deficiency is best treated by the intravenous route. Adults with good renal function should be given 12 mmol magnesium (in the form of magnesium sulphate in saline or 5% dextrose) over 2–3 h and a further 12 mmol infused slowly over the following 24 h. This regimen may need to be repeated over several days, as the tissue magnesium deficit takes longer to correct than the plasma concentration. If the plasma concentrations are increased too much, urinary magnesium losses increase. In infants, replacement is given in a similar manner, but the dosage should be reduced to 0.2–0.3 mmol/kg. Plasma magnesium concentrations need to be monitored during and after replacement therapy. The intramuscular route can also be used for replacement, but the injections are painful.

Where feasible, magnesium supplements may also be given orally (5–10 mmol/24 h, in divided doses). Suitable preparations include magnesium oxide and its sulphate and carbonate salts, which are available as antacids or purgatives. Higher doses may cause diarrhoea.

With drug-induced hypomagnesaemia, the offending agent should of course be withdrawn.

Hypermagnesaemia

The kidneys normally excrete excess magnesium very efficiently and can handle loads of up to 400 mmol/24 h, so hypermagnesaemia as a clinical problem is largely restricted to patients with acute or chronic renal failure. Hypermagnesaemia is, however, seen in familial hypocalcaemic hypercalcaemia (FHH), and can also occur with magnesium-containing enemas and cathartics. In subjects with renal failure, the potential sources of excess magnesium loads are the dialysate, magnesium-containing antacids (as phosphate-binding agents) and, in acute kidney injury, release of magnesium from tissues.

Hypermagnesaemia in patients with renal impairment can also occur with the use of citrate-gluconic acid solutions. These solutions, which contain substantial quantities of magnesium salts, are used either for bladder irrigation or are infused into the renal pelvis through a nephrostomy tube for the dissolution of renal stones.

Hypermagnesaemia in the order of 1.5–2.5 mmol/L may be associated with hypotension but is commonly asymptomatic. In the range 2.5–5.0 mmol/L, areflexia may be present and electrocardiographic changes occur (prolonged PR and QRS interval, peaked T waves). At higher concentrations, respiratory paralysis and cardiac arrest ensue: concentrations >8 mmol/L (almost always iatrogenic) are frequently fatal. At these very high plasma concentrations, magnesium has direct effects on neuromuscular cells; inhibiting acetylcholine release and causing peripheral blockade. Use is made of this in the pharmacological management of seizures associated with eclampsia of pregnancy, where an effective remedy is the parenteral administration of magnesium sulphate to maintain plasma concentrations at 3–4 mmol/L.

The treatment of hypermagnesaemia initially consists of cessation of magnesium administration. In an emergency, intravenous calcium gluconate (10 mL, 10% solution; 2.2 mmol calcium) will reverse the effects of hypermagnesaemia. Dialysis is also effective.

CONCLUSION

The physiology and pathology of calcium, phosphate and magnesium are closely related. The principal hormones involved in calcium and phosphate homeostasis are 1,25-dihydroxyvitamin D, FGF23 and parathyroid hormone. The great majority of calcium and phosphate is present in the skeleton, as is over half the magnesium, but all three ions have many other vital functions.

Pathological increases and decreases in concentrations of calcium, phosphate and magnesium can occur, and all can have serious consequences. For each disorder, a relatively small number of conditions account for the majority of presentations.

Further reading

Rosen CJ, editor. *Primer on the metabolic bone diseases and disorders of mineral metabolism*. 7th ed. Washington: American Society of Bone and Mineral Research; 2008.

A succinct and authoritative account of the current state of knowledge in the fields of bone biology, calcium metabolism and metabolic bone disease, written by the leaders in the respective fields.

Walton RJ, Bijvoet OL. Nomogram for derivation of renal threshold phosphate concentration. *Lancet* 1975;309–10.

APPENDIX 6.1: CALCIUM ABSORPTION TEST

This test was originally described by Broadus et al. (1978) for use in the investigation of hypercalcaemia, but it can also be used to estimate intestinal calcium absorption in other contexts. The patient takes a low calcium diet (10 mmol (400 mg)/24 h) for one week and is studied after an overnight fast (distilled water only). A water intake of 250 mL/h is maintained from 1 h before the test until its completion. The test consists of three consecutive 2 h urine collections for measurement of calcium and creatinine, with a plasma calcium and creatinine measurement at the midpoint of the first and third of these

collections. At the beginning of the second urine collection, a 1 g (25 mmol) oral calcium load is taken.

Interpretation

Reference values:

$$\begin{aligned} \text{rise in urine calcium output} &= 0.18 - 0.73 \text{ mmol} / 2 \text{ h} \\ \text{rise in plasma calcium} &= 0.07 - 0.30 \text{ mmol} / \text{L} \end{aligned}$$

Reference

Broadus AE, Dominguez M, Bartter FC. Pathophysiological studies in idiopathic hypercalciuria: use of an oral calcium tolerance test to characterize distinctive hypercalciuric subgroups. *J Clin Endocrinol Metab* 1978;47:751-760.

APPENDIX 6.2: ANALYSIS OF TUBULAR HANDLING OF CALCIUM

The relationship between urine calcium excretion per litre of glomerular filtrate (Ca_E) and plasma calcium has been established from calcium infusion experiments in healthy subjects. The renal tubular handling of calcium can be assessed by reference to this normal relationship. Ca_E is estimated from:

$$\frac{\text{urine [calcium]} \times \text{plasma [creatinine]}}{\text{urine [creatinine]}}$$

obtained from the second void sample after rising in the morning, while still fasting. All these concentrations are expressed as mmol/L. The units of Ca_E are mmol/L glomerular filtrate.

Interpretation

Hypercalcaemic patients falling to the right of the reference range (Fig. 6.9) are excreting less calcium than

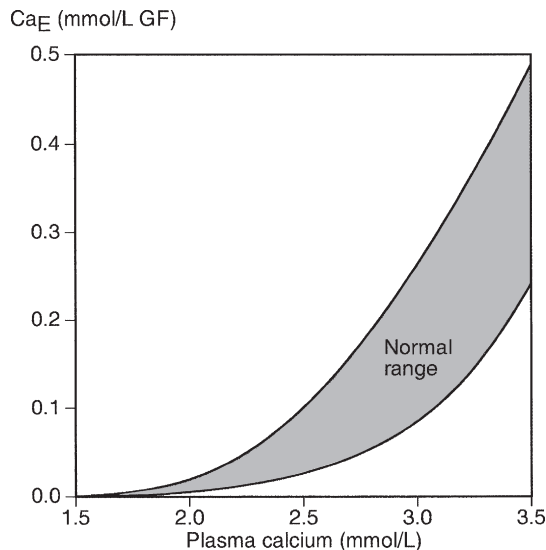


FIGURE 6.9 ■ Renal tubular handling of calcium (see text). (From Nordin B E C. Diagnostic procedures. In: Nordin B E C (ed). Calcium, phosphate and magnesium metabolism. Edinburgh: Churchill Livingstone; 1976, with permission).

expected for their prevailing plasma calcium concentration. Provided that this is documented at a time when the patient is not volume depleted, it can then be inferred that abnormally high renal tubular calcium reabsorption is contributing to the patient's hypercalcaemia. Typically, this is seen in primary hyperparathyroidism, familial hypocalciuric hypercalcaemia and malignant hypercalcaemia due to PTHrP. Similar considerations apply in the investigation of hypocalcaemia. Patients in whom Ca_E falls to the left of the reference range are excreting more calcium than would be predicted from the low plasma calcium concentration, and it can be inferred that abnormally low renal tubular calcium reabsorption is contributing to the hypocalcaemia (typically seen in hypoparathyroidism).

APPENDIX 6.3: CLASSIFICATION OF PSEUDOHYPOPARATHYROIDISM

This protocol is a modification of the Ellsworth-Howard and Chase-Auerbach tests. It measures the response of urine cAMP and TmP/GFR to an injection of synthetic human PTH (1-34).

The test is performed in a fasting patient, who ingests 250 mL/h of water between 06.00 and 12.00 noon. Two 30 min urine collections are collected before 09.00, at which time PTH (1-34), 0.625 µg/kg (maximum dose 25 µg) is infused intravenously over 15 min. Consecutive urine collections are made from 09.00-09.30; 09.30-10.00; 10.00-11.00 and 11.00-12.00, with blood samples taken at 09.00 and 11.00.

Analysis

Urine specimens: cAMP, creatinine, phosphate concentrations. Blood specimens: creatinine, phosphate concentrations. cAMP excretion is expressed as nmol/L glomerular filtrate (GF):

$$\frac{\text{urine [cAMP]} \times \text{plasma [creatinine]}}{\text{urine [creatinine]}}$$

Urine and plasma creatinine concentrations are expressed in the same units.

Phosphate excretion expressed as TmP/GFR (mmol/L glomerular filtrate; see Appendix 6.4).

Interpretation

Normal response:	10-20-fold increase in cAMP excretion, fall in TmP/GFR of 20-30%
Pseudohypoparathyroidism:	Type I, <5-fold increase in cAMP Type II, normal increase in cAMP Both types, <10% fall in TmP/GFR

Reference

Mallette LE, Kirkland JL, Gagel RF et al. Synthetic human parathyroid hormone (1-34) for the study of pseudohypoparathyroidism. *J Clin Endocrinol Metab* 1988;67:964-72.

APPENDIX 6.4: ESTIMATION OF TMP/GFR

A fasting, second voided urine sample is obtained in the morning. The convention is that this should be a 2 h collection, but the timing is in fact immaterial since time does not enter into the calculation. A plasma sample is obtained at the same time.

Measurements: plasma phosphate and creatinine; urine phosphate and creatinine (all expressed in the same units).

The ratio of the clearance of phosphate ($C_{\text{phosphate}}$) to the clearance of creatinine ($C_{\text{creatinine}}$) is calculated:

$$\frac{C_{\text{phosphate}}}{C_{\text{creatinine}}} = \frac{\text{urine [phosphate]} \times \text{plasma [creatinine]}}{\text{plasma [phosphate]} \times \text{urine [creatinine]}}$$

The tubular reabsorption of phosphate (TRP) = $1 - (C_{\text{phosphate}}/C_{\text{creatinine}})$.

TmP/GFR is read from the nomogram which relates the prevailing plasma phosphate concentration [PO_4] to TRP (Fig. 6.10).

A straight line through the appropriate values of plasma phosphate and TRP (or $C_{\text{phosphate}}/C_{\text{creatinine}}$) passes through the corresponding value of TmP/GFR. It should be noted that TmP/GFR and plasma phosphate concentrations are expressed in the same units. Two scales have been given: the 0.0–2.0 scale is suitable for estimating values of TmP/GFR close to the reference range expressed in SI units (0.80–1.35 mmol/L glomerular filtrate) and the 0.0–5.0 scale for values expressed in mass units (2.5–4.2 mg/100 mL glomerular filtrate).

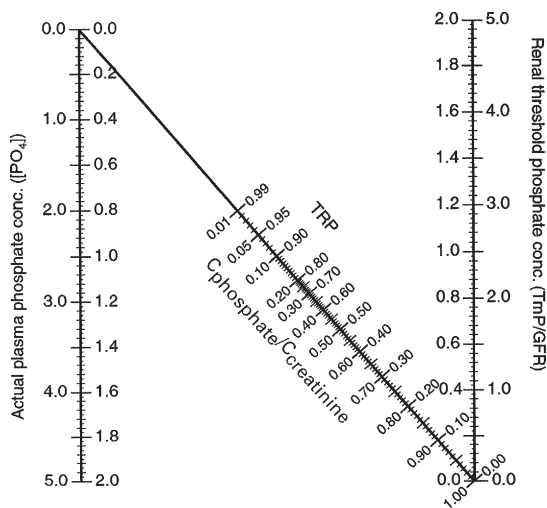


FIGURE 6.10 ■ Estimation of TmP/GFR (see text). (From Walton R J, Bijvoet O L M 1975 Lancet 2: 309–310, with permission).

APPENDIX 6.5: MAGNESIUM RETENTION TEST

Suggested schema for clinical use of magnesium tolerance test in normomagnesaemic patients:

1. Collect baseline urine (spot or timed) for magnesium/creatinine molar ratio
2. Infuse 0.1 mmol MgSO_4 /kg lean body weight in 50 mL dextrose saline over 4 h

3. Collect urine for 24 h (starting with infusion) for measurement of excretion of magnesium and creatinine (Cr) (both in mmol)
4. Calculate % Mg retained using the following formula:

$$1 - \left(\frac{\text{postinfusion urine [Mg]} - (\text{preinfusion urine [Mg]} - ([\text{Mg}]/[\text{Cr}] \times \text{postinfusion urine [Cr]})}{\text{total elemental Mg infused}} \right)$$

5. Criteria for Mg deficiency: >50% retention at 24 h.

Reference

From Ryzen E, Elbaum N, Singer FR, Rude RK. Parenteral magnesium tolerance testing in the evaluation of magnesium deficiency. *Magnesium* 1985;4:137–47. With permission of the publishers Karger, Basel.

APPENDIX 6.6: RENAL TUBULAR REABSORPTION OF MAGNESIUM

1. A fasting urine and blood sample are obtained after an overnight fast.
2. A loading dose of 0.4–0.8 mmol magnesium is infused (1–2 mL 10% MgSO_4 solution), and then a continuous infusion of 8 mmol over 2 h (in 50 mL 5% dextrose) is begun.
3. Over the third and fourth hours of the test, 16 mmol magnesium are infused.
4. For the fifth and sixth hours of the test, 32 mmol magnesium are infused. (Quantities shown are for an adult.)
5. Urine specimens are collected hourly for the measurement of magnesium and creatinine concentrations. Blood specimens are collected at the midpoint of each hour for the measurement of magnesium and creatinine.
6. A graph is plotted of magnesium excretion/litre of glomerular filtrate (Mg_{excr} : y-axis) vs ultrafilterable plasma magnesium (ufMg: x-axis).

$$\text{Mg}_{\text{excr}} = (\text{urine [magnesium]} \times \text{plasma [creatinine]}) / \text{urine [creatinine]}$$

$$\text{ufMg} = 0.75 \times \text{plasma [magnesium]}$$

The graph should show a line roughly parallel to the line of identity. The TmMg is the vertical distance between these parallel lines. A typical normal value is ~0.6 (meaning that when ufMg is less than 0.6 mmol/L, virtually all filtered magnesium is reabsorbed). In patients with impaired renal tubular reabsorption, TmMg values may be close to zero.

It is possible to combine both retention and tubular reabsorption tests in hypomagnesaemic patients, if the volume of each of the six urine samples is measured.

In addition to calculating magnesium excretion/litre of glomerular filtrate (as above) for each collection, the total amount of magnesium excreted over the 6 h of the test can be calculated. A collection is then made of the next 18 h urine to calculate 24 h urine excretion for determination of magnesium retention (as in Appendix 6.5).

The kidneys, renal function and kidney disease

David Makanjuola • Marta Lapsley

CHAPTER OUTLINE

ANATOMY 124

Gross anatomy 124

Microstructure 124

RENAL FUNCTION 127

Renal blood flow and its control 127

Glomerular function 127

Tubular function 128

RENAL DISEASE AND ITS PRESENTATION 129

Introduction 129

Manifestations of renal disease 129

Diseases affecting the kidneys 130

THE ASSESSMENT OF RENAL FUNCTION 130

Introduction 130

Biochemical tests of renal function 130

ACUTE KIDNEY INJURY (ACUTE RENAL FAILURE) 136

Introduction 136

Classification and causes 136

Prerenal acute kidney injury 137

Intrinsic acute kidney injury 138

Obstructive (postrenal) kidney injury 139

Acute kidney injury in the setting of chronic kidney disease 139

Hepatorenal syndrome 139

Metabolic consequences and management of acute kidney injury 140

CHRONIC KIDNEY DISEASE 141

Introduction 141

Aetiology and pathogenesis of chronic kidney disease 141

The uraemic syndrome 142

Growth retardation 144

Sexual dysfunction 145

Thyroid abnormalities 146

Anaemia 147

Endocrine control of salt and water balance 147

Carbohydrate metabolism and lipid metabolism 148

Management 148

CONCLUSION 151

Note about terminology 151

ANATOMY

Gross anatomy

The kidneys are located against the posterior abdominal wall on either side of the vertebral column. They lie anterior to the diaphragm and various muscles; their anterior surfaces are covered by parietal peritoneum. The left kidney is posterior to the stomach, pancreas, spleen and descending colon, and the right to the liver, the second part of the duodenum and the ascending colon. Their superior poles are covered by the suprarenal (adrenal) glands. In the adult, the kidneys are about 11–13 cm long, 6 cm wide and 4 cm thick. They weigh approximately 150 g each, yet together receive about 25% of the resting cardiac output. This blood is supplied by the renal arteries, which are branches of the aorta. Venous drainage is into the renal veins, which drain into the inferior vena cava. The kidneys are innervated by sympathetic nerves arising from the sympathetic chains and by parasympathetic fibres arising from the vagus nerve.

Each kidney is covered by a poorly distensible capsule. This limits the swelling that can occur during acute inflammation, and in such circumstances there is an increase in tissue pressure that tends to decrease the glomerular filtration rate (GFR).

Congenital absence of one kidney occurs in approximately 1 in 2400 individuals; this is only of significance if renal surgery is contemplated for any reason. The remaining kidney usually undergoes compensatory hypertrophy. Another common (approximately 1 in 10 000) congenital anomaly is the single ‘horseshoe’ kidney, in which the lower poles of the potential two kidneys are conjoined; renal function is normal but, occasionally, ureteric obstruction may predispose to calculus formation and infection.

Microstructure

Each kidney contains approximately one million functional units, called nephrons. These are tubular structures, consisting of various histologically and functionally

distinct elements. Each nephron has a single glomerulus, which is located in the outer part of the kidney (the cortex) and is responsible for the ultrafiltration of the blood. The remainder of the nephron consists of several contiguous tubular structures that progressively modify the composition of the ultrafiltrate before it is passed out as urine. These structures (Fig. 7.1) are:

- the proximal convoluted tubule, also located in the cortex
- the loop of Henle, which has the configuration of a hairpin and extends into the deeper part of the kidney (the medulla), then doubles back on itself to return to the cortex
- the distal convoluted tubule, located in the cortex
- the collecting duct, which extends down through the medulla to the renal papillae, whence the urine drains into the renal pelvis. The renal pelvises are drained by the ureters, and the right and left ureters themselves drain into the urinary bladder, where urine is stored prior to voiding.

The glomerulus

Each glomerulus consists of a tuft of capillaries that protrudes into the dilated, blind end of the nephron (the Bowman capsule) (Fig. 7.2). The nephron is composed of epithelial cells, and these and the endothelial cells of the glomerular capillaries are separated by a

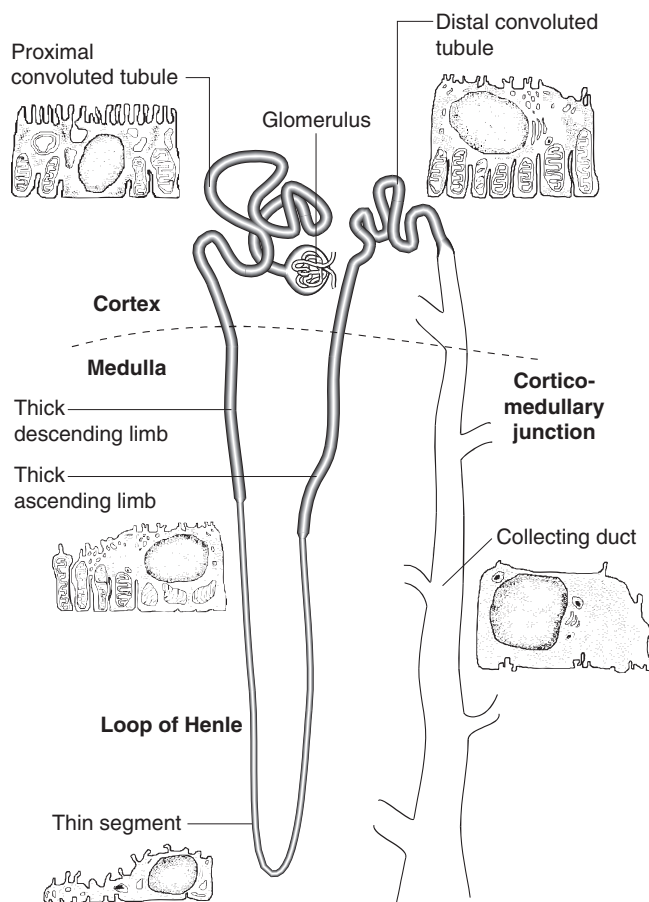


FIGURE 7.1 ■ Diagram of a nephron showing the structure of typical tubular cells.

basement membrane. This visceral basement membrane is continuous with a parietal basement membrane associated with the epithelial cells forming the outer part of the Bowman capsule, which are continuous with the epithelium of the proximal convoluted tubule. The microstructure of the glomerulus is described in detail in Chapter 8.

The glomerular filtrate is formed by the passage of fluid from the capillaries through fenestrations in the endothelial cells and through the basement membrane. This is covered on its epithelial aspect by the interdigitating foot processes of the epithelial cells, or podocytes, and the ultrafiltrate passes across a thin membrane (the epithelial slit diaphragm) into the filtration slits between the foot processes and then into the Bowman space, which is continuous with the lumen of the nephron. The selective retention of different constituents of the blood occurs at different stages. The endothelial fenestrations are too small to permit significant passage of erythrocytes, leukocytes or platelets; the major barrier to the filtration of proteins is mechanical and provided by the inner layer of the basement membrane, but the anionic surface layer also selectively retards anionic proteins. Finally, the foot processes and slit diaphragms also act as a mechanical barrier. The relative rate of progression of the constituents of the plasma into the Bowman space, therefore, depends on molecular mass, shape and charge. Glomeruli also contain mesangial cells. These stellate cells are contractile, are involved in the regulation of glomerular filtration and have phagocytic properties. The basement membrane does not extend between the mesangial cells and endothelial cells. This facilitates their ability to phagocytose large particles from the plasma. Their ability to ingest immune complexes plays an important part in the pathogenesis of certain forms of glomerular disease.

The proximal convoluted tubule

This structure is about 15 mm long and is composed of a single layer of inter-digitating epithelial cells united at their apices by tight junctions. The luminal surfaces of these cells have microvillous brush borders that provide the large surface area required for the absorptive function of the proximal tubule. The proximal tubule drains into a short straight segment directed towards the outer medulla and continuous with the descending limb of the loop of Henle.

The loop of Henle

The descending limb of the loop of Henle (Fig. 7.1) is composed of flat epithelial cells. The majority of the loops, whose glomeruli lie in the outer part of the cortex, are relatively short, but those with more deeply situated glomeruli have longer loops (up to 14 mm), which extend down into the medullary pyramids.

The thin descending limb turns back on itself, ascending towards the medulla, and the epithelial cells become cuboid, rich in mitochondria and have an invaginated luminal surface. The thick ascending limb extends towards the glomerulus of the same nephron and then leads into the distal convoluted tubule.

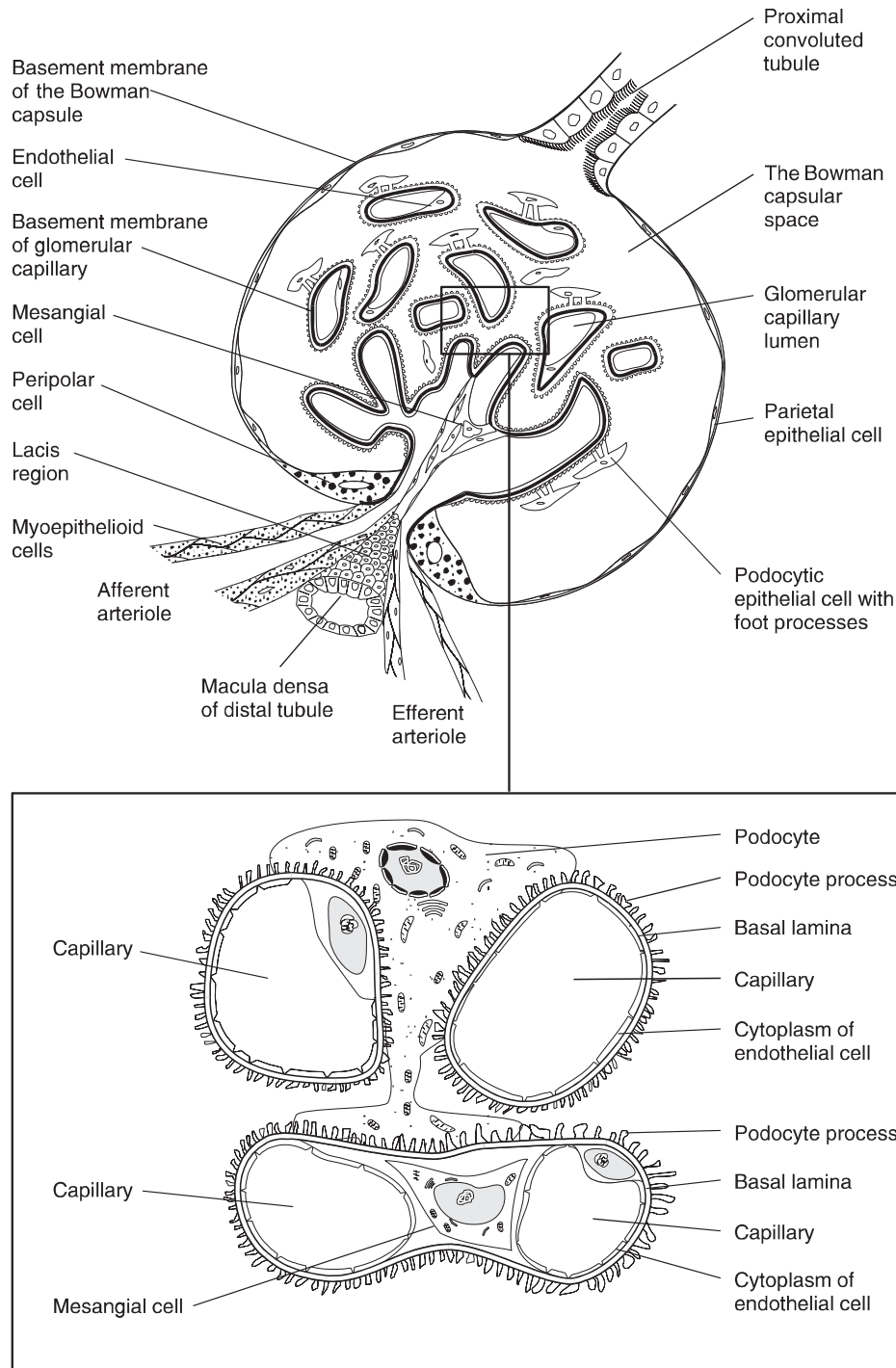


FIGURE 7.2 ■ Diagram of a glomerulus with detailed enlargement to show the cellular relationships.

The distal convoluted tubule and collecting duct

The distal convoluted tubule is about 5 mm long and is made up of typical absorptive epithelial cells, although without a brush border. It passes close to the afferent and efferent arterioles of the same nephron. The tubular epithelial cells in this area have a special function in relation to the control of renin secretion and are known as the macula densa. Renin is secreted by myoepithelioid juxtaglomerular cells in the walls of the afferent arterioles.

These cells, together with those of the macula densa and some other nearby cells (lacis cells), constitute the juxtaglomerular apparatus.

The distal convoluted tubule also contains intercalated (I) cells, responsible for the secretion of acid.

The collecting ducts are composed of I cells and also principal (P) cells, which become permeable to water under the influence of vasopressin. The ducts descend from the cortex down through the medulla, opening to the renal pelvis from the renal papillae.

Other specialized cells

In addition to the cells of the nephron and blood vessels, the kidneys contain medullary interstitial cells. These characteristically contain lipid droplets and secrete prostaglandins.

Blood vessels

Blood leaving glomerular capillary tufts passes not into venules, but into efferent arterioles. Those draining the cortical glomeruli divide into a capillary plexus that enmeshes the renal tubules. Those draining the juxtamedullary glomeruli form the vasa recta; these are bundles of vessels that plunge into the medulla to varying depths, forming capillary plexi that envelop the loops of Henle and then rejoin to form the ascending vasa recta, which eventually drain into the renal veins. The vasa recta form a countercurrent exchange system (see below), which is essential to the formation of both concentrated and dilute (with respect to plasma) urine.

RENAL FUNCTION

The principal functions of the kidneys are to excrete water-soluble (mostly nitrogenous) waste products and toxins from the body, and to control the volume and composition of the extracellular fluid (ECF). The kidneys also have important endocrine functions. It is perhaps surprising that the kidneys produce some 170 L of glomerular filtrate per 24 h, but only 1–2 L of urine. The bulk of the solvent and solute filtered is reabsorbed, but the control of the processes responsible gives considerable scope for the composition and volume of the urine to be adjusted to meet the physiological needs of the individual.

The physiology of the kidneys is well described in many standard physiology textbooks and will not be recounted in detail in this chapter. Rather, a brief summary is given and certain features that are pertinent to an understanding of the investigation of renal function and the mechanisms and consequences of renal impairment are emphasized.

Renal blood flow and its control

At rest, 20–25% of the cardiac output flows through the kidneys (1.1–1.3 L/min). The bulk of this blood perfuses the cortex; indeed, the flow rate to the medulla is actually less than to most other tissues.

The process of glomerular filtration is essential to renal function and is dependent on an adequate renal perfusion pressure. Autoregulatory mechanisms allow maintenance of renal blood flow and GFR within narrow limits in the face of a wide variety of external variables, including arterial pressure, venous pressure, ureteric hydrostatic pressure and plasma oncotic pressure. Renal blood flow and GFR are independent of mean arterial blood pressure over the range 80–200 mmHg. This is achieved by intrinsic, and humorally and neurally mediated, alterations in the tone of the renal blood vessels and, in particular, of the efferent glomerular arterioles.

Renal hypoperfusion stimulates the release of renin from juxtaglomerular cells. This enzyme converts circulating angiotensinogen to angiotensin I, which is in turn converted by angiotensin-converting-enzyme to angiotensin II, a powerful vasoconstrictor. This contributes to the maintenance of systemic blood pressure and, in the kidney, by causing efferent arteriolar vasoconstriction, helps to maintain the intraglomerular pressure despite a reduction in perfusion pressure. Various other mechanisms are involved in the maintenance of systemic blood pressure, but in the poorly perfused kidney, release of prostaglandins causes vasodilatation which, at least in mild hypotension, helps to maintain renal blood flow. A perfusion pressure of at least 50–60 mmHg is required to overcome the combined hydrostatic and oncotic pressures that oppose filtration; if mean arterial pressure falls below 80 mmHg, renal blood flow and GFR decline rapidly.

Glomerular function

Glomerular filtration is fundamental to the production of urine and to the homeostatic functions of the kidney. Techniques for measuring the GFR, the amount of filtrate formed per unit time, are described in a later section (p. 131).

The glomerular filtration rate is determined by the balance of pressures across the filtration barrier in the glomerulus, and the physical nature and extent of the barrier itself. The forces include the difference between afferent and efferent glomerular arteriolar pressures (promoting filtration) opposed by the difference in osmotic pressures between the ultrafiltrate and the plasma, and the hydrostatic pressure in the Bowman space. The net filtration pressure is normally approximately 15 mmHg at the afferent end of the capillaries, falling to zero at the efferent ends. Factors influencing the GFR are summarized in [Box 7.1](#).

A decrease in the area available for filtration can occur in many renal diseases in which there is glomerular damage. It can also result from contraction of mesangial cells mediated by agents such as angiotensin II, vasopressin, noradrenaline (norepinephrine), thromboxane A₂ and prostaglandin F₂; dopamine, natriuretic peptides and prostaglandin E₂ have opposite effects. A reduction in filtration area may be an important physiological mechanism for reducing GFR and thus conserving fluid.

BOX 7.1 Factors affecting glomerular filtration rate

- Renal blood flow
- Glomerular capillary hydrostatic pressure
 - Systemic blood pressure
 - Afferent/efferent arteriolar vasoconstriction
- Hydrostatic pressure in the Bowman space
 - Ureteric obstruction
 - Renal oedema (with restricting capsule)
- Plasma protein concentration
- Glomerular ultrafiltration coefficient
 - Glomerular permeability
 - Number of glomeruli

Tubular function

The proximal convoluted tubule

The proximal tubules are responsible for the active reabsorption of the bulk of filtered solute, accompanied by an iso-osmotic amount of water. Thus, virtually all the filtered glucose, amino acids, bicarbonate and potassium are absorbed here, together with some two-thirds of the filtered sodium. Absorption of solutes is an active, energy-requiring process and is isotonic, so that an equivalent amount of water is also absorbed and the fluid entering the descending limb of the loop of Henle is isotonic with plasma.

Transport occurs by way of ion channels, exchangers, co-transporters and pumps; filtered proteins are taken up into tubular cells by endocytosis and broken down into their constituent amino acids. Transport mechanisms have finite capacities and these may be exceeded in some circumstances. For example, glucose begins to be excreted in the urine if the plasma concentration exceeds about 10 mmol/L. Isolated or generalized disorders of renal tubular function result in the excretion of solutes normally reabsorbed in the proximal tubule even at normal plasma concentrations, as discussed further in Chapter 9.

The sodium content of the ECF is the primary determinant of ECF volume, the maintenance of which is critical to life. Given that about 26000 mmol of sodium are filtered at the glomeruli each day (equivalent to some 8–9 times the total body sodium), it is clearly essential that most of the filtered sodium is reabsorbed. Indeed, the bulk of sodium reabsorption is obligatory. However, an increase in GFR (which potentially could lead to a massive increase in sodium excretion) results in increased proximal tubular sodium absorption and *vice versa*. This process, termed *glomerulo-tubular balance*, is thought to be mediated largely through changes in the oncotic pressure in the peri-tubular capillaries. This increases when the GFR increases, owing to the increase in the volume of water removed from the plasma. The re-absorption of several other solutes is affected similarly. Normal kidneys are capable of excreting between virtually zero and more than 400 mmol of sodium per 24h. The fine control of sodium excretion is achieved in the distal convoluted tubule, where sodium reabsorption is stimulated by aldosterone.

Most of the filtered bicarbonate is absorbed in the proximal convoluted tubule, although indirectly. Hydrogen ions are generated in the tubular cells and are secreted into the tubular lumen (mainly in exchange for sodium) where they combine with the filtered bicarbonate ions. Bicarbonate ions are formed *pari passu* with hydrogen ions and are co-transported with sodium across the basolateral membranes of the tubular cells into the interstitial space (see Chapter 5).

Some substances are secreted into the urine from the proximal tubule. Examples include penicillins, creatinine, certain steroids and their glucuronides and derivatives of hippuric acid, for example *p*-aminohippuric acid (PAH). This property of PAH allows physiologists to use measurements of PAH clearance to determine renal plasma flow, but this is not an investigation that is used clinically.

The loop of Henle

This structure provides the countercurrent multiplier that generates the medullary hypertonicity that is essential for the regulation of water excretion and the production of concentrated urine.

Essentially, the process involves the active transport of solute (principally chloride and sodium ions) out of the thick ascending limb of the loop of Henle. This solute is not accompanied by water. As a result, the fluid within the lumen becomes hypotonic and the interstitial fluid surrounding the loop becomes hypertonic. Since water can diffuse out of the thin descending limb, but not the thick ascending limb, the net effect is that the fluid within the loop of Henle and the surrounding interstitial fluid become progressively hypertonic from the corticomedullary junction down into the medulla. There is further sodium and chloride absorption from the thick ascending limb. As water cannot diffuse out of the thick limb, the fluid entering the distal convoluted tubule is hypotonic with respect to plasma.

The anatomical disposition of the vasa recta allows them to act as a countercurrent exchanger, maintaining (though not actively contributing to) the osmotic gradient. The flows of water and solutes in the nephron are summarized in Figure 7.3. Whereas 60–70% of the filtered loads of solute and water have been reabsorbed from the glomerular filtrate at the beginning of the loop of Henle, a further 15% of water is removed during passage through this structure, so that only 10–20% of filtered water and somewhat less solute reach the distal convoluted tubule.

Tubuloglomerular feedback. There is a reciprocal relationship between the rate of flow of fluid through the ascending limb of the loop of Henle and the first part of the distal convoluted tubule of individual nephrons, and the rate of glomerular filtration through the same nephron. Thus, a decrease in flow rate increases filtration and vice versa. This process tends to result in a constant load of solute being presented to the distal convoluted tubule.

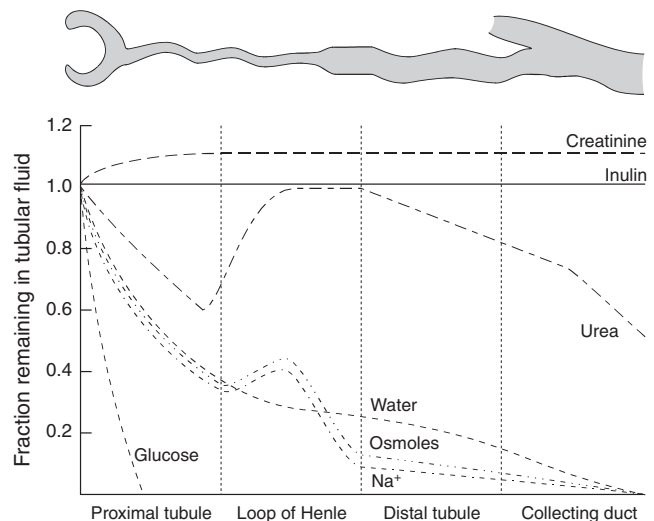


FIGURE 7.3 ■ Changes in the fraction of the filtered amounts of substances remaining in the tubular fluid along the length of the nephron.

Feedback is mediated through constriction and dilatation of the afferent arteriole by factors such as renal prostaglandins, adenosine and the renin–angiotensin system, in response to changes in chloride absorption in the macula densa. This mechanism becomes more sensitive when ECF volume is decreased and vice versa.

Tubuloglomerular feedback should not be confused with glomerulotubular balance, the change in proximal tubular sodium reabsorption in response to changes in GFR that has been discussed above and in Chapter 4.

The role of urea. Urea has an essential role in the generation of medullary hypertonicity. The thick ascending limb of the loop of Henle is permeable to urea (and sodium) but not to water. Movement of urea down its concentration gradient into the interstitium contributes significantly to medullary hypertonicity; the inner medullary portion of the collecting duct (see below) is also permeable to urea. The poor concentrating ability of the kidneys in the newborn is due to the decreased availability of urea to maintain medullary hypertonicity. In adults, renal concentrating power is dependent on an adequate protein intake (the source of urea) and is lower on a low protein diet.

The distal convoluted tubule

This part of the nephron is responsible for the ‘fine tuning’ of urinary composition. Aldosterone stimulates sodium reabsorption, generating an electrochemical gradient that allows the secretion (and thus excretion) of potassium and hydrogen ions. Normal distal tubular function is essential to the maintenance of hydrogen ion homeostasis.

The collecting duct

These structures extend from the ends of the distal convoluted tubules to the renal papillae. Their main function is to permit the reabsorption of solute-free water and thus regulate the osmolality of the urine. This is achieved through the action of vasopressin (antidiuretic hormone). The cells of the collecting ducts are normally impermeable to water. Under the influence of vasopressin, aquaporin-2 water channels are inserted into the tubular wall, such that they become permeable, allowing water to move out of the hypotonic fluid in the tubular lumen into the hypertonic interstitium, so producing concentrated urine. In the absence of vasopressin, no water is removed and the urine is hypotonic.

The extremes of achievable urine osmolality are 30–1400 mmol/kg. When maximally dilute, as much as 13% of the filtered water is excreted and the rate of urine production is about 16 mL/min; when maximally concentrated, <0.5% of filtered water is excreted and urine production is as little as 500 mL/24 h.

Some solute is also absorbed in the collecting ducts. In the cortical portion, sodium reabsorption takes place in exchange for potassium and hydrogen ions, as in the distal convoluted tubules. The medullary segment is partially permeable to urea, which moves passively out into the interstitium and helps to maintain the high osmolality of the medulla.

Diuresis. While the maximal urine flow rate when a water load is being excreted is approximately 16 mL/min (assuming an average solute load), higher rates are attainable if there are increased quantities of non-reabsorbed solutes in the tubules. This osmotic diuresis is due to the direct osmotic effect of the solutes and to a secondary effect on sodium reabsorption. The retention of water in the lumen of the proximal tubules increases the concentration gradient against which sodium must be reabsorbed; the same factor limits sodium reabsorption in the thick ascending part of the loop of Henle, thus interfering with the generation of medullary hypertonicity. This is the explanation for the diuresis characteristic of hyperglycaemia.

RENAL DISEASE AND ITS PRESENTATION

Introduction

There are many specific renal diseases; these may present with clinical features clearly referable to the kidneys (e.g. a decrease or increase in urine output), but frequently have systemic manifestations (e.g. the ‘uraemic syndrome’ (see p. 142) and hypertension). The kidneys can also be involved in multisystem disease, for example diabetes mellitus, the connective tissue disorders and amyloidosis.

Manifestations of renal disease

There are relatively few cardinal manifestations of renal disease and none is specific to any one underlying disorder. They include anuria, oliguria and polyuria; respectively a complete absence, decreased production and excessive production of urine. Polyuria may come to notice only by virtue of causing excessive production of urine at night (nocturia). These are usually symptoms of disordered renal function, although anuria and oliguria can be secondary to obstruction of the urinary tract and polyuria can be due to extrarenal disease that impairs the ability of the kidneys to concentrate the urine.

Urinary frequency (the passing of urine more frequently than normal, but in normal total 24 h volumes) must be distinguished from polyuria; it is often related to irritation of the urinary tract, for example by infection, but can also be due to prostatic enlargement. Irritation of the urinary tract is frequently accompanied by dysuria – pain on micturition. Patients with renal disease may experience loin pain or have tenderness in the loins. Renal or ureteric colic, an intermittent pain of considerable severity, is characteristic of the passage of urinary calculi.

Patients with renal disease may present with systemic abnormalities, including hypertension, pyrexia, oedema and the ‘uraemic syndrome’. This latter term is used to describe the protean manifestations of renal failure. The urine may appear abnormal in renal disease; the causes are discussed in a later section (p. 130).

Finally, renal disease may be revealed for the first time by the finding of a biochemical abnormality, particularly an elevated plasma creatinine or urea concentration, or positive urine dip-stick tests for protein or blood.

Diseases affecting the kidneys

Although patients with renal disease frequently present with an obvious functional abnormality, for example increased plasma creatinine or proteinuria, these are not specific to any one renal disease.

A full discussion of the nature of individual renal diseases is beyond the scope of this chapter. Many renal diseases specifically affect one part of the kidney. Glomerulonephritis, which is often of immunological origin, can manifest by decreases in the GFR or changes in glomerular permeability (e.g. causing proteinuria). Glomerulonephritis may be of primary renal origin, or occur secondarily to an extrarenal disorder (see Chapter 8).

A considerable number of conditions, both inherited and acquired, limited to the kidneys and having extrarenal manifestations, can cause disorders of renal tubular function. They are discussed in detail in Chapter 9.

The kidneys can be affected by infection, tumours and infiltrative and degenerative disorders. Generalized disorders, particularly shock, hypertension, connective tissue diseases and diabetes, are also frequent causes of renal dysfunction. The kidneys are also susceptible to the effects of toxins, including many widely used drugs (e.g. the aminoglycoside antibiotics and some cytotoxic drugs), as well as industrial and other toxins, and exposure to such agents is another frequent cause of renal dysfunction.

The major part of this chapter is devoted to a discussion of generalized renal failure, both acute and chronic. These are potentially life-threatening conditions and the clinical biochemistry laboratory plays a major part in their diagnosis and management.

THE ASSESSMENT OF RENAL FUNCTION

Introduction

Many techniques are of value in the investigation of the kidneys and renal function. The major groupings of techniques are laboratory tests, particularly biochemical and immunological, radiological imaging and the histological examination of renal tissue usually obtained by percutaneous biopsy.

Imaging techniques: standard radiography, ultrasonography, computed tomography, magnetic resonance imaging, the use of radionuclides and duplex scanning, can provide valuable anatomical information. Such techniques can be used to diagnose, for example, structural abnormalities, tumours, hydronephrosis, polycystic disease and the scarring of pyelonephritis. Radionuclide techniques can also provide information concerning renal function, particularly divided function (the separate function of the two kidneys) and glomerular function. Duplex scanning is a good screening test to assess renal blood flow before proceeding to magnetic resonance angiography or renal angiography.

Useful immunological markers include the detection in serum of antiglomerular basement membrane antibodies (positive in Goodpasture syndrome), antineutrophil cytoplasmic antibodies (in vasculitis), antinuclear antibodies and double-stranded DNA antibodies (in systemic

lupus erythematosus), complement components (in various types of glomerulonephritis) and antistreptolysin O antibodies (in post-streptococcal glomerulonephritis). Immunological techniques are also widely used in the histological examination of renal tissue. However, given the pivotal role of the kidneys in body fluid homeostasis, it is not surprising that biochemical investigations that indicate derangements of homeostasis are the most widely used tests of renal function. The remainder of this section covers such tests.

Biochemical tests of renal function

Urinalysis

The assessment of renal function should start with examination of the urine. Simple observations on the urine still contribute to the investigation of patients with suspected renal disease. To be valid, they must be made on a freshly passed sample.

Appearance. The normal colour of urine is due to urochrome pigments, and the deepness of the colouration, which can vary from a pale straw colour to deep amber, depends on urinary concentration. There are many causes of abnormal colouration of the urine, some of the more frequently encountered of which are indicated in [Box 7.2](#). In some cases, the colour is pH dependent. The distinction between haemoglobin or myoglobin and other causes of a reddish-brown urine can simply be accomplished by testing the urine with a suitable reagent stick; both myoglobin and haemoglobin will give a positive reaction for 'blood'.

Turbidity in a fresh sample suggests infection, but can also be due to the presence of fat in patients with the nephrotic syndrome. The presence of chyle in the urine may cause turbidity so severe that the urine appears milky. Anything more than a scant amount of foaming

BOX 7.2 Some causes of abnormal urinary colouration

Blue-green

- Methylene blue
- *Pseudomonas* infection
- Riboflavin

Pink-orange-red

- Haemoglobin
- Myoglobin
- Phenolphthalein
- Porphyrins
- Rifampicin

Red-brown-black

- Haemoglobin
- Myoglobin
- Red blood cells
- Homogentisic acid
- Levodopa
- Melanin
- Methyl dopa

when the urine is shaken, suggests proteinuria. The most frequent abnormal odour is that of ammonia, owing to the presence of urea-splitting microbes, either because the patient has a urinary infection or, in stale samples, because of contamination.

Specific gravity and osmolality. Both of these give an indication of the concentration of the urine. When this needs to be known accurately, for example in tests of renal concentrating ability, measurements of osmolality are essential. A reagent stick assessment of specific gravity may sometimes help in the interpretation of a screening test for proteinuria, since false negative results can occur with very dilute urine and vice versa. However, these reagent sticks detect ionic species only, and underestimate the specific gravity if other substances, for example glucose, are present.

pH. Normal urine is acidic, except after meals. Measurements of urine pH are important in the investigation of renal stone formers and as part of a test of urinary acidification in the investigation of suspected renal tubular acidosis.

Glucose. Glycosuria can occur either because of an increase in blood glucose concentration or because of a low renal threshold for glucose, owing to impaired proximal tubular reabsorption. In the latter instance, the blood glucose concentration will be normal. Renal glycosuria is discussed further in Chapter 9.

Protein. Proteinuria is an important feature of renal disease. Normal urine protein excretion is <200 mg/24 h and is not detectable with standard reagent sticks (e.g. 'Albustix'). Pathological proteinuria may be:

- overflow (due to raised plasma concentrations of low molecular weight proteins)
- glomerular (due to increased glomerular permeability to protein)
- tubular (due to decreased tubular reabsorption of filtered protein or saturated reabsorption)
- secreted (derived from the epithelium of the urinary tract).

Further details of the causes, detection and investigation of proteinuria are discussed in Chapter 8 and will not be considered further here.

Urinary sediment. Microscopic examination of the sediment obtained by centrifuging a freshly passed sample of urine frequently reveals a few cells (e.g. erythrocytes, polymorphonuclear leukocytes and cells derived from the kidney and urinary tract), hyaline (or rarely granular) casts composed of uromodulin (Tamm-Horsfall protein), and sometimes fat droplets, mucus threads and pigment granules.

Increases in any of these may occur in renal diseases. Abnormal casts are also frequently present. These are accretions of uromodulin with other material, for example cells or cellular debris. Red cell casts imply haematuria from glomerular disease, and white cell casts, the presence of white cells within the renal tubules. In acute pyelonephritis, there is an increase in polymorphs and various types of cast, in contrast to lower urinary tract infection where

polymorphs, but not casts, are increased. Microbiological examination in acute pyelonephritis will usually reveal bacteriuria. In acute glomerulonephritis, haematuria causes the urine to appear reddish-brown or smoky, and the sediment contains large numbers of red and white blood cells, and red cell, haemoglobin and other casts. In chronic glomerulonephritis, the amount of sediment is much less, but there is usually an excess of dysmorphic red blood cells, which suggests that red cells have passed through the glomerular membrane. Red cell and haemoglobin casts are sometimes present. Granular casts are seen in the urinary sediment in many renal diseases, but particularly in acute tubular necrosis; in this condition, casts containing tubular cells are also often present, but red cell casts are uncommon. Characteristic crystals may be seen when specific substances are present in excess or the pH of the urine promotes crystallization, for example, hexagonal cystine crystals, birefringent calcium oxalate crystals, calcium phosphate or uric acid crystals (see Chapter 9).

Other substances. Many other substances can be detected in the urine in appropriate circumstances, for example drugs or bilirubin. Their presence is due to the normal excretory function of the kidneys and is not a reflection of renal dysfunction, so they will not be considered further here.

Measurement of glomerular filtration rate

Glomerular filtration is essential to renal function, and investigations designed to measure the GFR are the most frequently performed tests of renal function. The factors that determine the GFR have been discussed above. Its measurement is based on the concept of clearance – the determination of the volume of plasma from which a substance is completely removed (the plasma is 'cleared' of the substance) by glomerular filtration during its passage through the kidney. Clearance is a theoretical concept – no substance is cleared completely from the plasma in this way – but can nevertheless be used as a valid measure of the GFR.

For the clearance of a substance to equal the GFR, it must be freely filtered by the glomeruli and eliminated from the body solely by this route (e.g. it must not be metabolized by the liver or secreted into or reabsorbed from the urine). The substance should be non-toxic and accurate measurement should be easily available in a routine laboratory. The clearance of a substance requires measurements of the plasma concentration and urinary excretion rate:

$$\text{Clearance} = (U \times \dot{V}) / P$$

where U is the concentration of the substance in the urine, \dot{V} is the rate of formation of urine and P the plasma concentration of the substance. It should be noted that \dot{V} is a rate; it has the dimensions (volume/time) and its determination requires the collection and measurement of the total amount of urine formed over a known period of time. The units of the various quantities are usually adjusted so that the clearance is expressed as mL/min.

TABLE 7.1 Properties of substances used to assess glomerular filtration rate

Property ^a	Urea	Creatinine	Inulin	^{99m} Tc-DTPA
Not protein bound	Yes	Yes	Yes	Yes
Freely filtered by glomerulus	Yes	Yes	Yes	Yes
Neither secreted nor absorbed in nephron	Flow-related reabsorption	Some secretion	Yes	Yes
Produced endogenously at constant rate	No	Yes	No	No
Easily measured	Yes	Yes	No	No

^aThe ideal substance would have all these properties. DTPA, diethylenetriaminepentaacetic acid.

The critical properties of some of the substances used in clearance measurements to assess GFR are shown in Table 7.1.

Clearance can also be used to determine renal plasma flow, using a substance such as *p*-aminohippuric acid, which is almost completely cleared from the blood in a single passage through the kidneys by a combination of glomerular filtration and secretion.

Inulin clearance. Inulin is a plant polysaccharide that satisfies all the physiological criteria mentioned in the previous section. The measurement of inulin clearance remains the 'gold standard' for the estimation of GFR. However, it is a relatively complex procedure and, although test kits are commercially available, it is infrequently used in clinical practice. In essence, the test involves the injection of a bolus dose of inulin, followed by a maintenance infusion designed to produce a constant plasma concentration. Once this has been achieved, a series of timed urine samples is collected and blood is drawn for the measurement of inulin at the midpoints of the collection periods. The GFR is taken as the mean of the inulin clearances for each period.

Creatinine clearance. Creatinine is an endogenous substance, a normal product of muscle metabolism. Its rate of production is fairly constant from day-to-day, being determined by muscle bulk rather than by activity. Creatinine is removed from the body mainly by glomerular filtration and its clearance can be measured as an index of GFR, although in the UK, it is much less frequently used than previously.

Creatinine is actively secreted into the urine, so that its clearance tends to overestimate the GFR. This effect is of little significance at normal filtration rates, the ratio (creatinine clearance/inulin clearance) being between 1.1 and 1.2, but in advanced renal disease the contribution of active secretion to the total amount of creatinine excreted in the urine becomes high in relation to the amount filtered, and creatinine clearance may significantly overestimate the true GFR. This is despite the fact that in established renal failure, bacterial degradation of creatinine secreted into the gut may contribute approximately 2 mL/min to the total clearance.

Inaccuracies arising from methodological problems with creatinine measurement have largely been overcome, but the major cause of inaccuracy in the determination of GFR by creatinine clearance is the accurate measurement of urine volume. Traditionally, patients have been required to collect urine over a 24h period. This is a

convenient time, but requires the patient to void their urine completely at the beginning of the 24h period and collect all the urine passed over the ensuing 24h, making a final collection at the end of this period. The possibilities for mistakes resulting in an incomplete collection are considerable and even under ideal conditions, for example in a metabolic unit with highly motivated patients, the coefficient of variation for repeated measurements in the same individual is >10%. The critical difference (the amount by which two estimates must differ to give a 95% probability that there has been a true change in GFR) is therefore >33%. The accuracy of estimates of the GFR by measurement of creatinine clearance may be improved by making two or more consecutive 24h urine collections, but this is often not practical even if it is acceptable to the patient.

There is, however, nothing special about the 24h period. The use of shorter periods, although more convenient for the patient, may reduce accuracy through inadequate bladder emptying, but a good compromise is to make a collection overnight. As long as the bladder was emptied at the beginning of the test (before retiring) and when the final collection was made (on rising), the urine production rate can be calculated and substituted in the $(U \times V)/P$ formula.

It is essential that a reliable method is employed to measure plasma and urine creatinine concentrations. Colorimetric assays tend to overestimate creatinine since they detect non-creatinine chromogens. But even when reliable methods are used, it should be appreciated that the creatinine clearance is based on four measurements: plasma and urine creatinine concentrations, urine volume and time. Each has an inherent inaccuracy and the overall analytical variance will be the sum of the individual variances.

Plasma creatinine concentration. The way in which creatinine is handled by the kidney, coupled with the fact that, in any one individual, the rate of production is relatively constant from day-to-day, means that the plasma creatinine concentration alone can be used as an index of renal function. But, although it is widely used, it has a number of disadvantages and these must be borne in mind when interpreting plasma creatinine concentrations. Perhaps the most important becomes apparent from considering the familiar $(U \times V)/P$ formula for deriving the GFR. Plasma creatinine concentration is *inversely* related to GFR, not *directly* related. This means that a plot of plasma creatinine concentration against GFR has the form of a hyperbola (Fig. 7.4). At any concentration of creatinine,

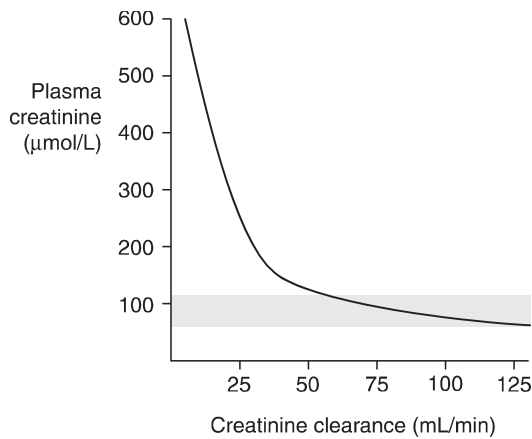


FIGURE 7.4 ■ Relationship between creatinine clearance and plasma creatinine concentration. As the clearance is related to the reciprocal of the plasma concentration, considerable loss of renal function can occur without the plasma creatinine concentration rising above the reference range (shaded).

halving the GFR will double the plasma creatinine concentration. But given the wide reference range for plasma creatinine concentration in the healthy population, this means that in an individual, the plasma creatinine concentration could be within the reference limits, yet the GFR be only one-half of normal.

This is of critical importance to clinical practice, since in early renal disease (when it might be supposed that therapeutic intervention would be most beneficial), the plasma creatinine concentration may be ‘normal’ – that is, within the reference range – in spite of the GFR being reduced. Plasma creatinine concentration is thus a relatively insensitive index of mild renal functional impairment. At low values of GFR, however, it becomes extremely sensitive, although its accuracy declines because of the increased contribution of tubular secretion to overall renal excretion.

Other factors that must be taken into account in assessing the significance of plasma creatinine concentrations include muscle mass, diet (a recent meal including meat, particularly if stewed, may cause a transient increase in plasma creatinine concentration) and the presence in the plasma of substances that interfere with the assay. Ketones, bilirubin, cephalosporins and spironolactone have all been implicated, depending on the method used, but especially in Jaffé-(picric acid) based assay systems. Enzymatic assays for creatinine are said to be more specific than older colorimetric methods, but they are still prone to interference by certain drugs such as acetylcysteine and phenindione.

The importance of muscle mass is exemplified by the changes that occur with age. In healthy individuals, GFR, and thus creatinine clearance, declines with increasing age from about the end of the 4th decade, at a rate of approximately 1 mL/min per year. However, plasma creatinine concentration does not normally rise with age; this is because of a decrease in production rate, thought to reflect the tendency for muscle mass to fall with age.

Although comparison of a measured plasma creatinine concentration with a reference range can be misleading, the fact that *intraindividual* variation is less than *interindividual*

variation means that the detection of a *change* in creatinine concentration in an individual provides a more sensitive indication of a change in renal function. Nevertheless, the biological and analytical variance (even using the best assays) results in a critical difference of about 17% (20 μmol/L at a plasma creatinine concentration of 120 μmol/L).

In patients with advancing renal disease, the progressive loss of renal function with time can be expressed as a graph of reciprocal plasma creatinine concentration against time (since GFR is proportional to $1/[\text{plasma creatinine}]$). A steady loss of renal function will cause this plot to be rectilinear (Fig. 7.5). Such plots are useful in patients with irreversibly declining renal function to help predict when renal replacement will become necessary, so that, for example appropriate means for vascular access for haemodialysis can be provided.

Given the problems associated with the measurement of plasma creatinine concentration and creatinine clearance, the practical application of these measurements to patients with, or suspected of having, renal disease is not straightforward. Since the rate of creatinine production is relatively constant in most individuals (the essential assumption being that lean body mass is constant), one approach is to measure the GFR (preferably by a reliable technique, see below) and the plasma creatinine concentration simultaneously and then to follow progress with serial plasma creatinine measurements.

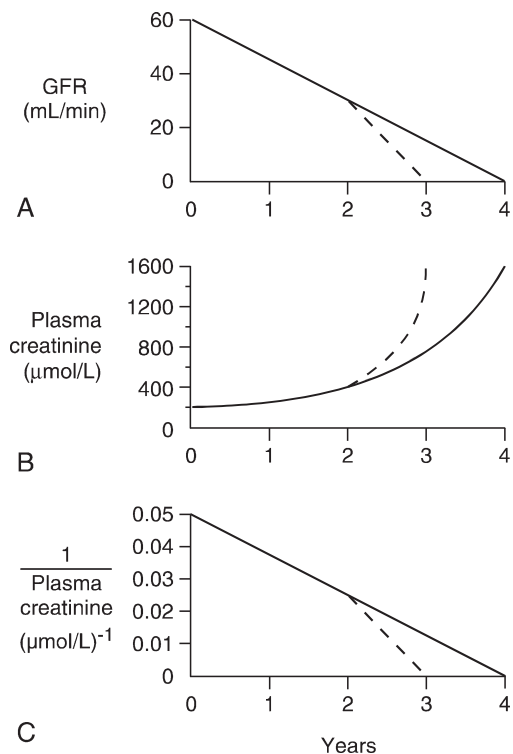


FIGURE 7.5 ■ The progression of chronic kidney disease. Hypothetical curves to show (A) the decline in glomerular filtration rate (GFR); (B) the increase in plasma creatinine concentration and (C) the decrease in reciprocal plasma creatinine concentration. The broken lines show the changes that might be expected if the decline in function was to accelerate for any reason (see Box 7.6).

Calculated creatinine clearance and estimated glomerular filtration rate. Given that the creatinine cleared by the kidney is almost entirely derived from muscle, and that muscle mass is related to body mass and to age, it is possible to derive formulae to estimate the creatinine clearance, and hence the GFR, from the plasma creatinine concentration. More than 25 different formulae have been suggested to correct for age, weight, gender and ethnic origin. The empirically derived equation of Cockcroft and Gault is one of the earliest and the best known of these. However, this and all similar formulae depend on measurements of plasma creatinine, and their accuracy can be impaired by all the factors that affect this. The early formulae, including that of Cockcroft and Gault, were derived using Jaffé based methods for the measurement of creatinine. Laboratories should take into account the relationship between their creatinine assay and the original assay used to derive any equation before advising clinicians on interpretation. None of the correction equations performs well in the physiological range of GFR, which limits their usefulness, for example in monitoring patients with diabetes or assessing potential kidney donors.

Cockcroft and Gault. The Cockcroft and Gault formula for males is:

$$C_{cr} = [1.224 \{ (140 - \text{age}) \times \text{wt} \} / C_{r_p}] \text{ mL/min}$$

where C_{cr} is creatinine clearance, C_{r_p} is plasma creatinine concentration in $\mu\text{mol/L}$, age is measured in years and wt is body weight in kilograms. For females, C_{cr} is 15% less than for males, so that the figure 1.224 in the above formula is replaced by 1.04.

Modifications to the formula are required before it can be used in children or obese adults as it assumes that increasing weight is related to increasing muscle rather than fat mass. The formulae developed for children are particularly useful, since there are considerable practical difficulties in making quantitative urine collections in this age group. The commonest estimation used in the UK is the Schwartz formula:

$$\text{GFR} = 49 \times \text{length} / C_{r_p} \text{ mL/min} / 1.73 \text{ m}^2$$

where length is measured in centimetres and plasma creatinine in $\mu\text{mol/L}$.

MDRD. The UK has adopted the system originally devised in the USA for deriving an estimate of GFR (eGFR), which can be used for population screening for kidney disease. This is based on the abbreviated MDRD (Modification of Diet in Renal Diseases) study equation. It is a useful estimate of GFR in adults up to a value of 60 mL/min per 1.73 m² body surface area. The equation was developed from measurements in an independent sample of 1070 individuals. Over 90% of the estimates were within 30% of the measured GFR (using ¹²⁵I-iothalamate), with only 2% having an error of more than 50%.

There are four variables (age, sex, race and creatinine concentration), but the equation only requires one biochemical value. The equation is being validated in other

patient groups, including patients with type 1 diabetes without microalbuminuria, those with mild decreases in renal function or a normal GFR, and non-US populations, for example Indian and Chinese.

The original MDRD equation included an additional two variables (albumin and urea), but there is very little advantage in using it, as any small gain in accuracy is offset by the increase in variability from the extra measurements.

The abbreviated MDRD formula for Caucasians is:

$$\text{eGFR (mL/min/1.73 m}^2) = 175 \times \{ (\text{plasma [Cr] in } \mu\text{mol/L}) / 88.4 \}^{-1.154} \times (\text{age in years})^{-0.203}$$

The factor 88.4 is required to convert $\mu\text{mol/L}$ to mg/dL (the units for which the formula was derived). This equation applies to males; for females, the result is multiplied by 0.742; for African-Caribbeans, it is multiplied by 1.212.

The original abbreviated MDRD formula had 186 as the multiplier rather than 175; the change was made to align the formula with measurements of creatinine by methods whose calibration is traceable to the reference method (isotope dilution with mass spectrometry).

In a study of serum creatinine measurements in laboratories, the method group mean bias for a specimen with an assigned value of 79.7 $\mu\text{mol/L}$ varied from -5.25 to +27.4 $\mu\text{mol/L}$. In order to standardize values for eGFR, UKNEQAS (UK National External Quality Assessment Service) has calculated additional corrections (slope and intercept adjusters) to align individual manufacturers' methods to the reference method.

It should be noted that the MDRD equation for eGFR has not been validated for use in children, acute kidney injury, pregnancy, oedematous states, muscle wasting diseases, amputees or malnourished individuals. It is less accurate in obese than in lean individuals. When the eGFR is >60 mL/min, many laboratories report the value as >60 mL/min per 1.73 m², owing to decreasing precision and accuracy at higher values of GFR. Values above 90 mL/min should never be reported as an absolute figure, because the correlation with measured GFR is poor.

CKD-EPI. The Chronic Kidney Disease Epidemiology Collaboration group produced a set of formulae based on pooled data from several studies, which correlate better with measured GFR than the original MDRD formula, especially at values >60 mL/min per 1.73 m². However, GFR is overestimated in certain groups of patients, especially those with a low body mass index. Different formulae are applied depending on the value of the serum creatinine. Further details are available on the websites of various renal organizations.

Definition of CKD using eGFR. A persistent reduction in eGFR (<60 mL/min) is defined as chronic kidney disease (CKD). However, patients with markers of kidney damage, like persistent albuminuria or abnormalities on imaging (such as polycystic kidneys) or on renal biopsy, are defined as having CKD even if their eGFR is normal.

The stage of CKD can be assigned based on the degree of renal GFR reduction, irrespective of diagnosis, according to the Kidney/Disease Outcomes Quality Initiative (K/DOQI) classification (see Table 7.2), and used to guide the appropriate management.

Cystatin C. This is a low molecular weight (13.4kDa) protein, a cysteine protease inhibitor that is present on the surface of all nucleated cells and from which it is shed into the plasma and excreted by glomerular filtration. Its plasma concentration is much less dependent than that of creatinine on weight and height, muscle mass, age or sex, and many studies have suggested that its plasma concentration is a more sensitive index of mild renal impairment than that of creatinine. Estimates of GFR based on either cystatin C measurements alone or combined with creatinine, are more accurate than those based on creatinine alone. Plasma cystatin C concentration may be increased in malignancy, hyperthyroidism and by treatment with corticosteroids (a disadvantage in kidney transplant recipients), independently of the GFR. However, its greatest drawback is that it has to be measured by immunoassay, which is much more expensive than the usual colorimetric or enzymatic techniques used to measure creatinine. Although not widely available in routine laboratories in the UK, measurement may have a role in the detection of early renal impairment in patients with extremes of muscle bulk, such as body builders and small elderly women and in postpubertal teenage boys, for whom children's reference ranges for creatinine are inappropriate.

Plasma urea concentration. Although of great historical importance and still widely used as a test of renal function, the measurement of plasma urea concentration suffers from many disadvantages for this purpose. Although the plasma urea concentration can provide useful information in some circumstances, particularly when it can be compared with the plasma creatinine concentration, the latter, in spite of the problems discussed above, is the more reliable test of renal function.

Urea is the end-product of the metabolism of many nitrogenous substances, particularly amino acids. It is freely filtered at the glomerulus, but its plasma concentration is dependent in part on its rate of formation, which can vary

widely according to the overall rate of protein turnover and can also be affected by hepatic function. For example, plasma urea concentration frequently rises following a gastrointestinal bleed, owing to increased urea formation from the digestion of the blood in the gut, irrespective of any effect of the blood loss on GFR, while patients with combined hepatic and renal failure may have normal plasma urea concentrations because of decreased production despite decreased excretion. There is significant passive reabsorption of urea from the lumen of the nephron, and this increases at high plasma urea concentrations (as in renal failure) and if the rate of flow of fluid through the nephron is low (as when GFR is decreased by dehydration). The familiar observation that the plasma urea concentration tends to rise before the creatinine in patients with 'prerenal' renal impairment (see p. 137) reflects this latter fact.

Plasma β_2 -microglobulin. This protein, a component of the major histocompatibility (HLA) complex, is shed into the plasma at a constant rate and, having a molecular weight of only 11 815 Da, passes freely through the glomeruli. As a result, its plasma concentration is normally very low (<2 mg/L), but it rises if renal function is impaired and can reach 40 mg/L. There is a direct relationship between plasma β_2 -microglobulin concentration and GFR, but its measurement has not been widely adopted for estimating the GFR for methodological reasons. Further, there are a number of conditions in which the production of β_2 -microglobulin is increased (e.g. lymphoid tumours and some inflammatory diseases) and these can result in increased plasma concentrations that do not reflect decreased clearance.

Isotopic techniques for measuring glomerular filtration rate. A number of radiolabelled compounds that are excreted entirely or largely by glomerular filtration have been used in tests designed to measure the GFR. These include ^{51}Cr -labelled ethylenediaminetetraacetic acid (EDTA), ^{125}I -iodothalamate and $^{99\text{m}}\text{Tc}$ -diethylenetriaminepentaacetic acid (DTPA). The Cr-EDTA-derived GFR is the accepted isotopic gold standard and is used when a very accurate estimate of GFR is required, for example in the work-up of potential kidney donors. In the

TABLE 7.2 Stages of chronic kidney disease

Stage	Description	eGFR ^a (mL/min per 1.73 m ²)	Action ^b
1	Kidney damage with normal or \uparrow GFR	≥ 90	Diagnosis and treatment, treatment of comorbid conditions, slowing progression, cardiovascular disease risk reduction
2	Kidney damage with mild \downarrow GFR	60–89	Estimating progression
3a	Moderate \downarrow GFR	45–59	Evaluating and treating complications
3b		30–44	
4	Severe \downarrow GFR	15–29	Preparation for renal replacement therapy
5	Established renal failure	<15 (or dialysis)	Replacement therapy (if uraemia present)

The suffix 'p' is used for all stages of CKD to denote the presence of proteinuria confirmed by laboratory testing.

^aeGFR is the estimated GFR.

^bIncludes action from preceding stages. Table based on the National Kidney Foundation K/DOQI chronic kidney disease classification with modifications recommended by UK NICE guidelines.

simplest type of test, a single bolus injection is given and accurately timed blood samples are collected over 2 h (longer may be necessary in patients with renal impairment). The logarithm of plasma radioactivity is plotted against time and extrapolated back to zero time, to allow calculation of the notional initial volume of distribution. The GFR is then given by:

$$\text{GFR} = V_0 \times (\log_e 2) t_{1/2}$$

where V_0 is the notional initial volume of distribution and $t_{1/2}$ is the half-life for the decrease in plasma radioactivity.

Such methods compare reasonably well with the results of inulin clearance, but their accuracy is enhanced if classic clearance techniques are used. For this, the marker compound is given during a water diuresis and a series of measurements of plasma and urine radioactivity are made. The clearance formula is used to calculate individual clearances and the GFR is taken as the mean of these.

^{99m}Tc -DTPA is also used as a renal scanning agent (although the dose of radioactivity required is approximately ten times greater than for measurement of GFR). However, gamma camera scanning following the administration of this agent can be used to measure the individual contribution of each kidney to the overall GFR.

Other tests of renal function

Tests of proximal tubular function include the measurement of phosphate reabsorption (see Chapter 6) and the detection of aminoaciduria and glycosuria. The causes of glycosuria are discussed in Chapters 9 and 15. Amino aciduria can be due to increases in plasma amino acid concentration ('overflow amino aciduria'), as occurs in several inherited disorders of amino acid metabolism, or to decreased renal tubular reabsorption of normally filtered amounts of amino acids ('renal amino aciduria'). Renal amino aciduria can be due to isolated defects of amino acid transport or to generalized proximal tubular dysfunction, and is discussed in Chapter 9.

Tests of distal tubular function include formal tests of urinary concentration and dilution and of urinary acidification. These are discussed in Chapters 4 and 9. The assessment of urinary sodium excretion in relation to physiological requirements plays an important part in the investigation of renal function in certain circumstances, as discussed in the section on acute kidney injury (see later). In general, however, tests of renal tubular function are employed much less frequently than tests of glomerular function in the investigation of renal disease.

ACUTE KIDNEY INJURY (ACUTE RENAL FAILURE)

Introduction

Acute kidney injury (AKI) has now replaced the term acute renal failure. The new terminology enables health-care professionals to consider the disease as a spectrum of injury. This spectrum extends from less severe forms of injury to more advanced injury requiring renal replacement therapy (RRT). Clinically, AKI is characterized by

a rapid reduction in kidney function resulting in a failure to maintain fluid, electrolyte and acid-base homeostasis. Acute kidney injury is potentially a life-threatening condition and, developing as it often does in patients who are already severely ill, it has a high mortality, despite the widespread availability of effective renal replacement therapy.

In many patients with AKI, there is oliguria (a urine flow rate of <15 mL/h in the adult). However, this is not universally present. In particular, non-oliguric AKI may be seen in patients with burns, liver disease and drug-induced renal damage. Occasionally, there is no urine production at all (anuria). This can occur with particularly severe renal disease but is more frequently due to urinary tract obstruction (to which it is an important clue).

It is important to appreciate that a diagnosis of AKI on its own is incomplete and does not imply any particular cause. The principal clinical features of AKI are summarized in Table 7.3.

The use of the adjective 'acute' tends to emphasize the importance of the period of time over which the condition develops. While it is true that this time is usually only a few hours (in contrast to that for chronic kidney disease, CKD), a more fundamental difference between these two syndromes is that AKI is usually potentially reversible, whereas in CKD, there is progressive, irreversible loss of renal function over a period that can range from months to tens of years, leading eventually to established renal failure, a condition in which renal replacement (e.g. long-term dialysis or transplantation) is required for continued survival.

Classification and causes

As has been discussed above, the primary determinants of the GFR are the hydrostatic pressure of blood in the glomerular capillaries, the intrinsic physical properties of the glomeruli and the hydrostatic pressure of fluid in the lumina of the nephrons. It follows that changes in

TABLE 7.3 Metabolic and clinical features of acute kidney injury

Metabolic feature	Clinical feature
Retention of nitrogenous waste products (↑ plasma urea and creatinine)	Nausea, vomiting, hiccoughs, clouding of consciousness, gastrointestinal bleeding etc.
Retention of sodium and water (usually with hyponatraemia)	Peripheral and pulmonary oedema, ascites, pleural effusion
Retention of potassium (with hyperkalaemia)	Typical electrocardiographic changes (peaked T waves, flattened P waves, QRS depression, widened QRS complex)
Retention of acid (metabolic acidosis, ↑ [H ⁺], ↓ pH, ↓ [HCO ₃ ⁻], ↓ PCO ₂)	Muscle weakness, paralysis, risk of cardiac arrest Kussmaul breathing, hypotension

any one or more of these may cause a decrease in GFR. Acute kidney injury has traditionally been classified into prerenal (due to impaired renal perfusion), intrinsic renal and postrenal (obstructive) causes (Box 7.3). This classification, though particularly useful in formulating an approach to diagnosis and management, has limitations; in any patient, more than one factor may be contributing to the development of renal failure. Furthermore, both 'prerenal' and 'postrenal' AKI can lead to the development of intrinsic renal damage. Also, intrinsic renal disease can have both detrimental effects on renal perfusion and lead to obstruction of urinary flow. Finally, the syndrome of acute kidney injury can develop in patients

BOX 7.3 Some causes of acute kidney injury

Prerenal

- Hypovolaemia
 - Haemorrhage
 - Gastrointestinal fluid loss
 - Diuresis
 - Burns
- Decreased cardiac output
 - Cardiogenic shock
 - Massive pulmonary embolus
 - Cardiac tamponade
- Other causes of hypotension
 - Sepsis
 - Vasodilator drugs
- Interference with autoregulation of renal blood flow
 - Angiotensin-converting-enzyme inhibitors
 - Prostaglandin inhibitors (e.g. non-steroidal anti-inflammatory drugs)

Intrinsic renal

- Glomerulonephritis
- Vascular disease
- Severe hypertension
- Ischaemia
- Nephrotoxins
 - Aminoglycoside antibiotics
 - Non-steroidal anti-inflammatory drugs
 - X-ray contrast media
 - Heavy metals
 - Animal and plant toxins
- Hypercalcaemia
- Infiltrative disorders
 - Sarcoidosis
 - Lymphoma, leukaemia

Postrenal

- Bilateral ureteric obstruction (or unilateral if there is only one functioning kidney)
 - Calculi
 - Papillary necrosis
 - Tumours
 - Retroperitoneal fibrosis
 - Surgical mishap
- Bladder outflow/urethral obstruction
 - Urethral stricture
 - Benign prostatic hypertrophy
 - Carcinoma of prostate
 - Neurogenic bladder

with CKD as when, for example, complete bladder outflow obstruction leads to retention of urine and AKI in a man who already has chronically impaired renal function as a result of partial obstruction due to prostatic hypertrophy.

Prerenal acute kidney injury

This type of injury is characterized, at least in its early stages, by a lack of structural renal damage, preservation of normal tubular function (e.g. concentrating ability, sodium reabsorption) and rapid reversibility provided that the underlying cause is managed appropriately. In essence, it is a consequence of reduced renal perfusion, secondary to either cardiovascular insufficiency (hypovolaemia and hypotension) or a derangement of intrarenal haemodynamics. Some causes are indicated in Box 7.3. In many hospitals, a significant proportion of cases of acute kidney injury follows trauma or surgery and may be exacerbated by the administration of nephrotoxic drugs or coexistent medical conditions.

In health, renal blood flow and GFR remain unchanged over a wide range of renal perfusion pressures, but a fall in perfusion pressure below a mean of about 80 mmHg leads to a reduction in renal blood flow and GFR.

Homeostatic mechanisms involved in the defence of ECF volume include increased sodium reabsorption and increased water retention. The major stimulus to sodium reabsorption is through angiotensin II-induced secretion of aldosterone. Water reabsorption is stimulated by increased vasopressin secretion; stimuli to vasopressin secretion in this context may include hypovolaemia, angiotensin II and, if hypovolaemia is due to hypotonic fluid loss, increased plasma osmolality. In addition, vasoconstriction of efferent glomerular arterioles by angiotensin II increases the filtered fraction of the plasma, thus increasing the protein content (and oncotic pressure) of blood in these vessels and in the peritubular capillaries, so enhancing proximal tubular reabsorption of water.

Diagnosis

The accurate diagnosis of prerenal AKI is important, since rapid intervention may prevent progression to intrinsic AKI. In prerenal AKI, the plasma urea concentration is usually increased disproportionately to that of creatinine. As urine flow rate falls, the diffusion of urea from the lumen of the nephron into the interstitial fluid increases. An increase in urea may also be related to the increased catabolism that frequently is present as a result of the underlying disorder. In contrast, in intrinsic kidney injury, plasma urea and creatinine concentrations tend to rise in parallel.

Urinalysis may be informative and should be performed in all patients with AKI. As discussed on page 130, abnormalities apart from the presence of hyaline casts are uncommon in prerenal AKI, but are invariably present in intrinsic AKI. The fact that tubular function is intact in prerenal AKI underlies a number of diagnostic tests (see Table 7.4). In prerenal AKI, the homeostatic responses to hypovolaemia described above should

TABLE 7.4 Diagnostic tests of potential value in differentiating prerenal and intrinsic renal acute kidney injury

Test	Result	
	Prerenal	Intrinsic renal
Urine [Na ⁺]	<20 mmol/L	>40 mmol/L
Urine osmolality/plasma osmolality	>1.5	<1.1

maximize renal sodium reabsorption and water retention. In consequence, the urine will tend to be concentrated and have a low sodium content, whereas if there is tubular damage, there will be a failure of this response, and the urine will be dilute and contain significant quantities of sodium. Although in general these theoretical conclusions are supported by clinical observations, considerable overlap may occur in the results obtained in patients with prerenal and established AKI. It must be emphasized that the tests are only useful if there is oliguria, and may be vitiated if the patient has been given a diuretic (which will increase sodium excretion, even in prerenal AKI), or mannitol or an X-ray contrast medium, both of which increase urine osmolality even in intrinsic AKI.

Although these tests are of help in management, the diagnosis of intrinsic AKI may only be made on the basis of an inexorably rising plasma creatinine concentration despite correction of any 'prerenal' factors that may have been identified.

Management

The management of prerenal AKI involves rapid restoration of euvoemia, discontinuation of potentially nephrotoxic drugs and increasing tubular flow to prevent tubular obstruction. Delay in achieving euvoemia increases the risk of progression to intrinsic kidney injury. Fluid replacement must be done with a view to achieving adequate cardiac output but avoiding fluid overload. In patients with shock, it may be necessary to infuse vasoactive substances to achieve an adequate blood pressure.

Intrinsic acute kidney injury

Intrinsic AKI can be a progression from a prerenal phase, but many conditions causing AKI do so without a prerenal component. Nephrotoxins, intrinsic renal disease (e.g. glomerulonephritis) and systemic disease affecting the kidneys (e.g. septicaemia) are all important causes of kidney injury (see [Box 7.3](#)).

The term 'acute tubular necrosis' is sometimes used synonymously with 'intrinsic AKI', but should not be. Ischaemic injury and many nephrotoxins cause acute tubular necrosis, but in other cases of renal injury, the brunt of the damage is borne by the glomeruli. Although the treatment in such patients will involve management of the kidney injury per se, the prognosis often depends on the underlying cause.

Diagnosis

The distinction between prerenal and intrinsic AKI has been discussed above. In addition, proteinuria is usually present in intrinsic injury. The urinary sediment frequently contains epithelial cells, both free and in casts, and in kidney injury due to glomerulonephritis, haematuria and red cell casts are present.

The plasma biochemical disturbances are similar in all types of kidney injury though without intervention, they become more severe if prerenal AKI progresses to intrinsic injury. The plasma urea and creatinine concentrations then tend to rise in parallel.

Special investigations that may be of value (but will not be required in all cases) include radiology (ultrasound, isotopic scanning, CT and MRI scans) and renal histology.

Ultrasound is helpful in the exclusion of urinary tract obstruction. It shows the size and cortical thickness of the kidneys (small kidneys imply chronicity). Asymmetry of renal size may suggest renovascular disease. This can be confirmed by duplex scanning. If poor images are obtained, blood flow can be determined by magnetic resonance angiography (MRA). This technique does not require the use of X-ray contrast media, but in patients with eGFR <30 mL/min, the gadolinium which is used as a contrast agent for MRI scanning has been implicated in the development of a condition known as nephrogenic systemic fibrosis. There is no known cure for this condition. Computed tomography (CT) scanning is also useful in further definition of anatomical renal tract abnormalities. Contrast-enhanced CT scanning can, however, exacerbate renal impairment by causing contrast nephropathy. It should, therefore, be used with caution in patients with acute kidney injury.

Renal biopsy is not required if the cause of AKI is obvious (e.g. post-trauma), but will inform management if the development of kidney injury is unexpected and if intrinsic renal disease, for example glomerulonephritis, is suspected.

Acute tubular necrosis

Pathogenesis. While the reduction in GFR that occurs with renal hypoperfusion is readily explicable, the cause of its persistence after restoration of the circulation is a more complex issue, which is still not fully understood. Several mechanisms may be involved ([Fig. 7.6](#)). These include continued renal vasoconstriction due to the intrarenal release of vasoactive substances, such as endothelin and prostaglandins, and to angiotensin II (renin secretion may remain high secondary to decreased delivery of solute, especially sodium, to the macula densa); direct damage to the glomeruli resulting in decreased filtration, and physical obstruction of the lumina of nephrons by swollen tubular cells or tubular debris. Diffusion of fluid from the lumina of nephrons into the interstitial tissue through damaged tubular walls (stimulated by the difference in oncotic pressures), will tend to oppose the process of glomerular filtration.

There is evidence from experimental models to support all these mechanisms. In all probability, all are

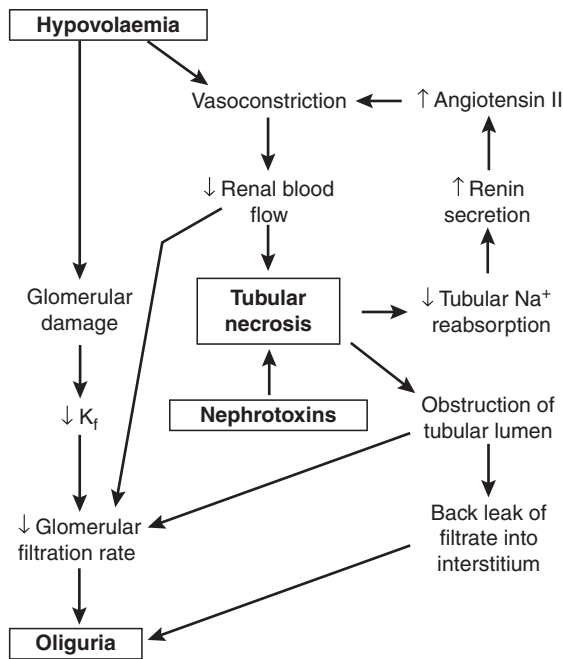


FIGURE 7.6 ■ Factors involved in the pathogenesis of oliguria in acute tubular necrosis. K_f , glomerular ultrafiltration coefficient.

involved in the pathogenesis of acute tubular necrosis, but the relative contributions of the various mechanisms are likely to differ in each case.

Whatever the cause, however, injury, whether nephrotoxic or ischaemic, appears to be central to the development of acute tubular necrosis, and the restoration of normal renal function requires not only restoration of glomerular filtration, but also regeneration of the tubular epithelium of nephrons.

Natural history. The natural history of acute tubular necrosis (ATN) can be divided into three phases: the initial, oliguric phase; a diuretic phase, as glomerular filtration begins to recover, but a combination of persisting tubular damage and a high osmotic load causes a diuresis, and a recovery phase. The time course is variable between a few days and several months. Although patients who survive usually recover normal renal function, a small number develop renal cortical necrosis, from which functional recovery is not possible.

Exceptions occur and some patients do not have an oliguric phase. The details of management of each phase are different although the principles – avoidance of complications and the provision of renal replacement therapy as required – are common to all patients with AKI.

The mortality of AKI has decreased considerably with the widespread availability of renal replacement treatment, but remains in the region of 50%. It is related not only to the severity of the kidney injury and the efficiency of treatment, but also to the underlying cause and any comorbidities, for example multiple trauma. Acute kidney injury frequently occurs in patients who are already severely ill. The prognosis is generally better in non-oliguric AKI than when oliguria is present.

Obstructive (postrenal) kidney injury

The possibility of an obstructive cause must be considered in any patient with AKI. The history and examination (e.g. previous symptoms suggestive of prostate disease, or a palpable bladder on clinical examination) may provide a clue. Ultrasonography may reveal dilatation of the ureters if there is ureteric obstruction.

Obstructive AKI may be due to urethral (e.g. benign prostatic hypertrophy), bilateral ureteric (e.g. calculi, retroperitoneal fibrosis) or intrarenal (e.g. light chain precipitation in myeloma) obstruction. Any of these conditions can also cause chronic kidney disease.

The procedures involved in the diagnosis of obstructive AKI often reveal the cause. Management is directed at restoring a normal flow of urine, for example by percutaneous nephrostomy, or removal or ultrasonic destruction of calculi. Relief of obstruction often leads to a temporary diuresis and natriuresis, and the urine output should be monitored closely to guide fluid replacement therapy.

Acute kidney injury in the setting of chronic kidney disease

Acute kidney injury can occur and precipitate presentation in patients with pre-existing (chronic) kidney disease. Clues to this possibility are indicated in [Box 7.4](#). In practice, it is often necessary to manage the acute event and investigate the patient for chronic renal disease later.

Hepatorenal syndrome

Patients with liver disease can develop acute kidney injury. The causes may be common (e.g. paracetamol (acetaminophen) toxicity), or independent. Prerenal and intrinsic renal disease can occur. The hepatorenal syndrome is a unique form of AKI that can develop in patients with chronic liver failure. It is discussed in detail in Chapter 14. The hepatorenal syndrome can progress to acute tubular necrosis; this has a particularly poor prognosis and recovery usually only occurs if there is some improvement in hepatic function.

BOX 7.4

Factors suggesting an element of chronicity in acute kidney injury (i.e. acute on chronic kidney disease)

- Absence of acute illness
- Long duration of symptoms
- Nocturia
- Anaemia
- Bone disease
- Sexual dysfunction
- Neuropathy
- Skin disorders
- Small kidneys demonstrated by radiological imaging

Metabolic consequences and management of acute kidney injury

The importance of prevention has already been emphasized. Many cases of AKI occur in patients at predictably high risk of developing the condition.

Accurate diagnosis is essential in the management of acute kidney injury. It is essential to consider the following possibilities:

- is there a prerenal component?
- is there any urinary tract obstruction?
- is there any intrinsic renal disease?
- is there a chronic component?

The approaches to answering these questions have been discussed above.

If renal function is not rapidly restored to normal despite appropriate management of any antecedent condition, the aim of treatment becomes the prevention of the consequences of the condition. There are two aspects to management: general measures and, if these alone are insufficient, renal replacement therapy. Specific measures to treat the underlying cause should, of course, be continued. All patients require rigorous clinical and laboratory monitoring throughout.

General management

Fluid and electrolyte balance. The prevention of fluid overload or electrolyte imbalance is crucial to the management of AKI. In calculating fluid balance, allowance must be made for endogenous water production of some 500 mL/24h and insensible losses (approximately 0.5 mL/kg body weight/h, but significantly higher if there is pyrexia). Endogenous production derives from oxidative metabolism and will be increased if (as is usual) the patient is catabolic.

Objective assessment of extracellular fluid (ECF) volume, preferably using flow-based measurements (or, alternatively, central venous pressure), is valuable, and in many patients is essential. Sodium intake must be limited in oliguric patients and special note taken of any covert sodium administration if drugs are given as their sodium salts. In patients with non-oliguric AKI and during the diuretic phase of acute tubular necrosis, adequate fluid replacement is required and should be informed by continuous assessment of urinary volume.

Hyperkalaemia is a life-threatening complication of AKI and frequent monitoring of plasma potassium concentration is essential. Several factors contribute to the increase in plasma potassium concentration, including decreased renal excretion, acidosis and loss of intracellular potassium to the ECF as a result of catabolism. Potassium intake should be minimized; intravenous calcium chloride can be used to antagonize the effects of hyperkalaemia. An infusion of dextrose and insulin may temporarily reduce the plasma potassium concentration, but persistent significant hyperkalaemia is an indication for renal replacement treatment.

Other disturbances that may occur in AKI include hyperphosphataemia, hypocalcaemia, hyperuricaemia and hypermagnesaemia. Hypercalcaemia may develop, particularly in patients with AKI related to crush injuries,

owing to release of calcium from damaged muscle and increased synthesis of calcitriol as falling phosphate concentrations decrease the inhibition of renal 1α -hydroxylase.

Acid–base balance. Since the kidneys are the only organs capable of excreting significant amounts of hydrogen ions, and the net daily acid load that has to be excreted is of the order of 40–80 mmol, patients with AKI are usually acidotic. Acid–base status may also be adversely affected by comorbid conditions, for example lactic acidosis secondary to sepsis.

Nutrition. Active nutritional management is a vital part of the management of patients with AKI, whether being treated conservatively or with renal replacement. The aim is to maintain nutritional status where it is adequate and to improve it where it is compromised, in order to limit catabolism, promote the healing of any wounds and enhance resistance to infection. The mainstay of nutritional management is to provide adequate energy substrates to obtain positive energy balance. Nitrogen requirements are variable and there are few reliable data to guide the amount of protein or amino acids that should be provided. In established renal failure, an increase in plasma urea concentration not associated with an increase in creatinine suggests excessive provision of nitrogen. It is important to provide adequate vitamins and trace elements.

If possible, the enteral route should be used to provide nutrients. In patients who cannot eat, or eat enough, proprietary liquid feeds may be given through a nasogastric tube. For patients with intestinal failure, and for those in whom enteral feeding is contraindicated for some other reason or in whom enteral feeding cannot, on its own, provide sufficient energy, parenteral feeding will be required.

The principles of nutritional management in AKI are no different from those in any other condition: an adequate regimen must be provided; a suitable route for its provision must be available, and the patient's response must be carefully monitored (see Chapter 11). It should be appreciated that patients receiving renal replacement in general have increased requirements for nutrients, since there will be loss from the bloodstream into the dialysate.

There is no reason not to use lipid emulsions together with glucose solutions as an energy source in patients with AKI receiving parenteral feeding, unless there is a problem with lipid utilization; this will be apparent from simple inspection of the plasma, which will appear lipaemic.

Other measures. Infection is a frequent cause of death in patients with AKI, even if it has not precipitated the condition. It is essential to treat any infection early and adequately and to take every practicable step possible to prevent infection developing. When prescribing drugs, including antibiotics, care must be taken to avoid any that are potentially nephrotoxic. The dosages of any drugs that are excreted by the kidneys should be reviewed and, where appropriate, dosage guided by measurement of plasma concentrations.

Renal replacement treatment

Major indications for renal replacement include the following:

- fluid overload (e.g. pulmonary oedema)
- the need to 'create space' to allow provision of parenteral nutrition or other solutions in the oliguric patient
- severe hyperkalaemia (plasma potassium >7.0 mmol/L)
- severe acidosis (arterial $[H^+]$ >70 nmol/L, pH <7.15 or plasma bicarbonate <12 mmol/L).

Haemodialysis. This technique is discussed on page 149. It is usually performed intermittently, each treatment lasting approximately 4 h. Clearance rates of more than 30 mL/min can be achieved. Intermittent haemodialysis is an efficient technique, but there is a danger of 'disequilibrium syndrome' due to too rapid removal of metabolites leading to osmotic imbalance between body fluid compartments with consequent rapid shifts of fluid. Headache, vomiting and, occasionally, fits may occur.

Because excess fluid is removed intermittently, the provision of continuous parenteral feeding may be more difficult with haemodialysis than with the filtration techniques described in the following sections.

Continuous venovenous haemofiltration (CVVH).

This technique (see p. 150) allows clearance rates of approximately 10 mL/min. In haemofiltration, large volumes of fluid are removed, but over a longer period of time, thus minimizing the risks of circulatory imbalance. There is, therefore, ample scope for the provision of parenteral nutrition when required, although other fluid replacement may usually be required in addition.

Continuous venovenous haemodiafiltration (CVVHDF).

In this technique, the filter is also perfused by dialysis solution. This combines the advantages of continuous filtration (particularly the ease of removal of fluid) with those of dialysis (higher clearances). Correction of metabolic imbalances, for example hyperkalaemia, is more easily achieved with CVVHDF than CVVH.

Peritoneal dialysis. Peritoneal dialysis (see p. 150) is a simpler technique than haemodialysis or haemofiltration, and does not require access to the circulation. However, it is less efficient than haemodialysis (clearance rates approximately 20 mL/min) and so is less suitable for hypercatabolic patients. Considerable quantities of protein (up to 40 g) can be lost into the dialysate at each session and there is a risk of peritonitis. Nevertheless, if the requirements needed to set up haemodialysis or filtration are not immediately available, this technique is valuable in the management of patients with AKI.

CHRONIC KIDNEY DISEASE

Introduction

Chronic kidney disease is a syndrome of persistent renal impairment involving loss of both glomerular and tubular function such that the kidneys' homeostatic functions

are compromised. It usually has a gradual onset and, in many cases, progresses inexorably to a critical state (established or end-stage renal failure) in which the patient's continued survival requires the initiation of renal replacement treatment, either by some form of dialysis or by renal transplantation. Although there are many causes of CKD, the clinical features of the condition tend to be similar regardless of the cause, being in effect due to a decrease in the number of functioning nephrons. Despite this, it is, as discussed below, important to attempt to determine the cause, although this is not always possible.

Although CKD may be discovered early, before it has given rise to any clinical disturbance (e.g. by the routine assessment of renal function in patients at risk of developing CKD), it frequently presents late, with a GFR already <20 mL/min. Some patients will also have clinical features of the 'uraemic syndrome'. This term is used to describe the clinical features that can occur in patients with CKD, but is caused by the retention of many substances, of which urea itself is relatively unimportant, and by other (e.g. endocrine) abnormalities that arise as a result of renal dysfunction.

Aetiology and pathogenesis of chronic kidney disease

Some of the more frequent causes of CKD are indicated in Box 7.5. As alluded to above, it is not always possible to determine the cause; patients presenting with advanced uraemia frequently have small, shrunken kidneys (unless diabetes, amyloid or polycystic disease is the cause of the condition) and specific diagnostic features may not be discernible even in a biopsy. It is, however, important to try to determine the cause, particularly in patients with less advanced disease. This may reveal a condition that is susceptible to treatment, so that progression can be delayed or even halted. The cause will, to some extent, determine the prognosis and allow planning of the eventual need for renal replacement and, if transplantation is contemplated, of the likelihood of recurrent disease in the grafted kidney. Finally, the identification of a familial cause (e.g. polycystic disease) may usefully lead to screening of other members of the family.

Successful treatment of the underlying cause of CKD does not always arrest the progress of the condition. It appears that, once started, the loss of nephrons in CKD itself contributes to disease progression. This has important therapeutic implications and the possible mechanisms have been extensively investigated.

BOX 7.5 Some causes of chronic kidney disease (CKD)

- Glomerulonephritis
- Pyelonephritis
- Interstitial nephritis
- Diabetes mellitus
- Hypertension
- Polycystic kidney disease
- Nephrotoxic drugs
- Multisystem disease

The progression of loss of renal function

If 80% of a rat's renal mass is removed, the animal develops proteinuria, hypertension and progressive renal failure; glomerular sclerosis is demonstrable in the renal remnant. Before hypertension and histological changes develop, there is a massive increase in renal blood flow, owing to a decrease in afferent arteriolar resistance. This leads to an increase in capillary hydrostatic pressure, increased capillary permeability and passage of macromolecules, including plasma proteins and lipoproteins, through the capillary wall. Some of the protein appears as proteinuria, but some is scavenged by mesangial cells, and it is thought that overloading of these phagocytic cells may cause functional derangement and cellular proliferation and contribute to the glomerular sclerosis. Increased numbers of macrophages may liberate growth factors, also stimulating cellular proliferation, and direct endothelial damage may lead to platelet activation and intraglomerular thrombosis, with consequent fibrosis and liberation of platelet-derived growth factor and other stimulants of cellular proliferation.

Systemic hypertension is common in chronic kidney disease (though less so if the underlying pathology is in the medulla, e.g. reflux nephropathy, than in the cortex, e.g. glomerulonephritis), owing to salt and water retention and inappropriate release of renin and, thus, formation of angiotensin I. This hypertension itself has a harmful effect on renal function, by exacerbating the processes described above and by causing arteriolar damage and ischaemic glomerular sclerosis.

Chronic kidney disease causes hyperphosphataemia, and the phosphate retention can contribute to further renal damage. Dyslipidaemia is a frequent finding in chronic renal disease, and the accumulation of lipids in the glomeruli is also thought to contribute to the progression of renal disease and suggests another possible therapeutic manoeuvre that may retard this process.

In addition to these factors, incidental conditions can contribute to the progression of renal failure in patients with renal disease (see [Box 7.6](#)). As the homeostatic capacity of the kidneys decreases, so the patient becomes more susceptible to what might otherwise be easily tolerated events.

BOX 7.6 Incidental factors which can contribute to the progression of CKD

- Hypertension
- Hypovolaemia
 - Vomiting, diarrhoea, excessive use of diuretics etc.
- Urinary tract infection
- Congestive cardiac failure
- Drugs
 - Tetracyclines
 - Non-steroidal anti-inflammatory agents
 - Angiotensin-converting-enzyme inhibitors
- Pregnancy
- Exacerbation of underlying disease (e.g. systemic lupus erythematosus)

The uraemic syndrome

The clinical syndrome associated with CKD is protean in its manifestations. None is specific to the condition, but the overall clinical disturbance is characteristic. Disorders of the cardiovascular, neurological, skeletal, gastrointestinal, endocrinological, haematological and immunological systems are of particular clinical importance, but any system can be affected. The clinical features are summarized in [Box 7.7](#).

BOX 7.7 Some clinical features of the uraemic syndrome

Cardiovascular

- Pericarditis
- Hypertension
- Congestive cardiac failure

Dermatological

- Pruritus
- Pigmentation
- Slow wound healing

Endocrinological

- Glucose intolerance
- Impaired growth
- Sexual dysfunction

Gastrointestinal

- Hiccough
- Nausea
- Vomiting
- Anorexia
- Gastritis

Haematological

- Anaemia
- Bleeding diathesis

Immunological

- Increased susceptibility to infection
- 'Burn out' of immunological disease (e.g. systemic lupus erythematosus)

Neurological

- Confusion, coma
- Dementia
- Headache
- Peripheral neuropathy
- Flapping tremor
- Muscle weakness

Pulmonary

- Pulmonary oedema

Skeletal

- Vertebral fractures

Other

- Nocturia
- Thirst
- General malaise

BOX 7.8 Uraemic toxins

- Urea
- Creatinine
- Oxalate
- Guanidines
 - Methylguanidine
 - Guanidinosuccinic acid
- Indoxyl sulphate
- *p*-Cresol and *p*-cresylsulphate
- Hippuric acids
 - Hippuric acid
 - *o*- and *p*-hydroxyhippuric acid
- Peptides
 - Parathyroid hormone
 - β_2 -microglobulin
- Polyamines
 - Spermine
 - Putrescine etc.

These changes result to a considerable extent from the effects of the retained 'uraemic toxins'; no single substance is responsible, although some, for example parathyroid hormone, appear to be particularly important. **Box 7.8** indicates some of the toxins that have been identified, but it should be emphasized that not only is it incomplete, but many uraemic toxins still await identification. In addition, deficiencies, for example of the hormones erythropoietin and calcitriol, contribute to the uraemic syndrome; while dialysis may relieve the clinical features of the syndrome by removing toxins, those related to deficiencies (e.g. anaemia) require treatment by specific supplementation.

Clinical features

A comprehensive discussion of the clinical manifestations of CKD is beyond the scope of this chapter. In the cardiovascular system, hypertension, congestive cardiac failure and accelerated atherosclerosis are all frequently encountered. The neurological involvement can include both the central nervous system (uraemic encephalopathy) and the peripheral nervous system. The skeletal features of CKD, which used to be called renal osteodystrophy, but is now called chronic kidney disease – mineral and bone disorder (CKD-MBD), comprise a complex amalgam of osteomalacia, secondary hyperparathyroidism, osteosclerosis and osteopenia; they largely stem from the impaired production of calcitriol (1,25-dihydroxycholecalciferol), the hormone derived from vitamin D. More recently, fibroblast growth factor 23 (FGF-23) and klotho, a transmembrane protein produced by the osteocyte, which is required for FGF-23 receptor activation, have been identified as being important in this area, but their exact roles are still under investigation. Renal bone disease is discussed in more detail in Chapter 31. Gastrointestinal problems include hiccoughs, anorexia and gastritis, which can all cause great distress to the patient; gastrointestinal bleeding can be life-threatening.

The anaemia of CKD is usually normochromic and normocytic (see p. 147); fragmented ('burr') cells may be present on the blood film. A multifactorial bleeding

diathesis is often present, with disturbances of platelet function and of coagulation. Depression of both the humoral and cellular arms of the immune system are frequent in CKD and patients have increased susceptibility to infection, are more likely to develop cancers and – a possibly beneficial effect – may experience remission of immunologically mediated disease.

Metabolic disturbances in CKD

Retention of nitrogenous waste products. Patients with CKD invariably have high plasma concentrations of urea and creatinine, but both can be affected by extrarenal factors. Thus, plasma urea concentration can be significantly reduced, even in advanced disease, by restriction of dietary protein, or in patients with chronic liver disease, and a sudden increase may occur if the patient suffers a gastrointestinal haemorrhage. Patients with CKD often have muscle wasting, which results in decreased creatinine production.

Hyperuricaemia is usually present in patients with CKD, although the concentration does not often exceed 600 $\mu\text{mol/L}$ and attacks of acute gout solely due to the CKD are rare.

Potassium metabolism. Potassium balance can often be maintained in CKD until the GFR falls to <5 mL/min. This is because of an adaptive response whereby distal tubular potassium secretion is increased; the mechanisms include increased aldosterone secretion and increased sodium delivery to the distal nephron, but there is also a poorly understood direct effect on potassium excretion. Furthermore, colonic potassium excretion, which is not usually a major route of potassium loss, becomes quantitatively important in patients with CKD.

Hyperkalaemia tends to occur earlier in patients with hyporeninaemic hypoaldosteronism, for example secondary to diabetes, and if there is significant acidosis. Hyperkalaemia is a life-threatening complication of CKD and can be precipitated by the injudicious administration of potassium or a drug that interferes with potassium excretion, for example potassium-sparing diuretics, angiotensin-converting-enzyme (ACE) inhibitors or β -adrenergic blockers.

Acid-base metabolism. In health, the kidneys excrete some 40–80 mmol of hydrogen ions per 24 h, this being the net rate of production on a normal diet. In CKD, renal hydrogen ion excretion is impaired, and once the GFR falls to below ~30 mL/min, systemic acidosis is likely to develop. This is despite the fact that the pH of the urine can usually be reduced to the lowest level achievable in health. The causes of the acidosis are discussed in detail in Chapter 5, but in summary are primarily due to reduced phosphate excretion, which diminishes the buffering capacity of the urine, and reduced ammoniogenesis which, albeit indirectly, also decreases renal hydrogen ion excretion.

In addition, there is often a partial defect in the reabsorption of filtered bicarbonate, a consequence, in part, of the expanded ECF volume and, in part, of the increased secretion of parathyroid hormone (see below). Severe

acidosis is, however, uncommon in CKD and, despite a continuing positive hydrogen ion balance, the blood hydrogen ion concentration tends to remain stable for weeks or even months at a time. This appears to be due to the extensive buffering of hydrogen ions that occurs in bone.

Although the usual acid–base disturbance in CKD is a non-respiratory acidosis, the ability of the kidneys to excrete excess bicarbonate is also decreased, and patients in whom there is excessive non-renal loss of acid, for example due to vomiting, may occasionally develop alkalosis.

Calcium, phosphate and magnesium metabolism and renal bone disease. These topics are reviewed in detail in Chapters 6 and 31. In essence, CKD leads to a progressive decrease in the synthesis of calcitriol from 25-hydroxycholecalciferol, as a result of both inhibition of the enzyme by retained phosphate (an effect that becomes apparent as the GFR falls to <60–70 mL/min) and, in more advanced disease, a decrease in the amount of enzyme as the renal mass decreases. A lack of calcitriol leads to decreased absorption of calcium from the gut, exacerbated by a low dietary intake of calcium, common in patients with CKD. There is thus a tendency to hypocalcaemia which, in health, should be corrected by increased secretion of parathyroid hormone (PTH) and mobilization of calcium from bone. In CKD, although the plasma PTH concentration is characteristically increased, often to a very high level (secondary hyperparathyroidism), there is resistance to its action, probably because of the low calcitriol concentrations. In consequence, plasma calcium concentration is usually below normal.

It should be noted that the results of older assays for PTH may be misleading, owing to metabolism of the hormone and selective retention of the biologically inactive C-terminal fragment in CKD. Assays that measure only intact PTH are more reliable indicators of parathyroid activity.

As discussed in detail in Chapter 6, hypercalcaemia can occur in CKD if parathyroid autonomy develops or after successful renal transplantation (tertiary hyperparathyroidism).

Plasma phosphate concentrations are usually normal, but can be low in early CKD. Despite a decrease in the number of functioning nephrons, the fractional excretion of phosphate by each is increased due to the elevated PTH. In more advanced disease, further loss of nephrons leads to the characteristic phosphate retention.

Plasma magnesium concentrations tend to be increased in CKD. Although toxic concentrations do not occur in uncomplicated cases, problems have been encountered in patients given magnesium salts therapeutically, for example as antacids.

Protein metabolism. Patients with CKD tend to be mildly catabolic, although this process can be accelerated by the development of intercurrent illness. Although the limitation of protein intake is an important aspect of the conservative management of CKD (see p. 148), it must not be so severe that it becomes a contributory factor to the loss of body protein.

Endocrine disturbances in CKD

There is good reason to expect endocrine disturbances in renal disease since the kidneys are an important site of hormone production, action and clearance (Table 7.5). Quite apart from the effect of the kidneys themselves on the production, excretion and activity of various hormones, impairment of non-endocrine renal function can secondarily affect the endocrine system. For example, CKD causes significant alterations in the internal environment that may influence the secretory control of hormones, their transport, activation and degradation, the relative amounts of free and protein-bound hormones in plasma and the responsiveness of hormone receptors at the cellular or subcellular level. Thus, in addition to renal clearance, the extrarenal clearance of some hormones may be abnormal in CKD, for example impaired hepatic degradation of biologically active PTH or diminished breakdown of insulin by skeletal muscle. Also, the function of endocrine organs may be influenced by treatment such as dietary restriction, drug therapy and renal replacement treatment.

The major mechanisms whereby chronic kidney disease influences endocrine function are outlined in Table 7.6. When kidney disease becomes advanced, the plasma or tissue concentrations and/or the functions of most hormones are altered, resulting in a number of well-recognized endocrine abnormalities (Box 7.9).

Growth retardation

Short stature is common among children with advanced kidney disease; its pathogenesis is multifactorial (Box 7.10). From an endocrine point of view, growth is a complex process involving primarily growth hormone (GH), but it is also dependent upon normal adrenal, thyroid and sex hormone function as well as an adequate supply of nutrients. In CKD, plasma immunoreactive GH concentrations rise in proportion to the decline in creatinine clearance and probably also to protein malnutrition.

TABLE 7.5 Functions of the kidneys involving hormones

Function	Hormones
Endocrine organ	Erythropoietin Angiotensin I and II 1,25-Dihydroxycholecalciferol Prostaglandins Kinins
Endocrine target organ	Parathyroid hormone Arginine vasopressin Aldosterone Calcitonin Atrial natriuretic peptide
Excretion of hormones	Cortisol Aldosterone Catecholamines Thyroid hormones Sex hormones
Catabolism of hormones	Parathyroid hormone Calcitonin Arginine vasopressin Insulin 1,25-Dihydroxycholecalciferol

TABLE 7.6 Mechanisms by which chronic kidney disease influences endocrine function

Mechanism	Example
Damage to renal parenchyma	Reduced synthesis of hormones Reduced clearance of hormones Reduced action of hormones
Disturbance of fluid and electrolyte balance	Activation of homeostatic mechanisms, e.g. renin–angiotensin system; PTH secretion in response to hypocalcaemia
Inappropriate secretion of hormones	Hypersecretion of hormones, e.g. prolactin
Appropriate homeostatic feedback mechanisms	Secretion of LH in response to reduced testosterone
Impaired extrarenal catabolism	Accumulation of hormones, e.g. PTH, insulin
Reduced hormonal function	Presence of multimolecular forms with varied biological activity, e.g. glucagon and calcitonin Accumulation of hormones with opposing actions Accumulation of metabolic inhibitors
Renal loss of binding proteins	Disturbance of bound:free ratio
Target organ resistance to hormonal action	Abnormal receptor binding, e.g. 1,25-dihydroxyvitamin D Abnormal postreceptor events, e.g. PTH, insulin
Reduced clearance of drugs	May interfere with endocrine function
Specific drug therapy, e.g. steroids	May interfere with endocrine function or induce endocrine disease, e.g. Cushing syndrome
Malnutrition	May have profound effects on endocrine function, e.g. growth and reproductive function
Increased susceptibility to sepsis	May have profound effects on endocrine function

BOX 7.9 Major endocrine abnormalities seen in CKD

- Anaemia
- Endocrine hypertension
- Glucose intolerance
- Growth retardation
- Hypergastrinaemia
- Hyperlipidaemia
- Renal osteodystrophy
- Sex hormone dysfunction
- Thyroid hormone abnormalities

BOX 7.10 Pathogenesis of growth retardation in CKD

- Malnutrition
- Anaemia
- Uraemia
- Renal tubular dysfunction
- Renal osteodystrophy
- Steroid therapy

This rise in GH is because of reduced clearance and consequently increased plasma half-life. Growth failure in the presence of raised plasma GH concentrations indicates a resistance to the actions of GH in peripheral tissues. The homeostatic mechanisms regulating GH secretion also appear to be abnormal in CKD, with exaggerated GH responses to arginine, thyrotrophin releasing hormone (TRH) and insulin-induced hypoglycaemia.

The skeletal growth-promoting actions of GH are mediated by the insulin-like growth factors (IGFs). In CKD, the plasma concentrations of IGFs tend to be normal, but their bioavailability is reduced. This is probably because of increases in the plasma concentrations of their respective binding proteins.

Sexual dysfunction

Impaired sexual function is a frequent and distressing symptom in both male and female patients with advanced CKD. In men, the major problem is impotence, which persists despite dialysis. This may have a multifactorial aetiology, being either directly related to the CKD or to coexistent conditions (Box 7.11). However, in many patients the organic component of impotence can be directly related to a disorder of sex hormones (Box 7.12), and several mechanisms may contribute to the hormonal abnormalities found. For example, hyperprolactinaemia is secondary both to hypersecretion and a reduction in clearance. The former is thought to be caused by a decrease in the sensitivity of lactotrophs to dopamine; chromatographic analysis of the circulating prolactin has confirmed that the majority is intact prolactin and not fragments.

In addition to problems of potency, males with CKD are frequently infertile, with evidence of defective spermatogenesis. Testicular biopsy shows maturation arrest, thickening of tubular basement membrane and interstitial fibrosis, consistent with damage to seminiferous tubules. These observations lend support to the theory that sexual dysfunction in advanced CKD is due primarily to toxic effects of circulating 'uraemic toxins' on both Leydig

BOX 7.11 Factors that may contribute to impotence in CKD

- Uraemic neuropathy
- Vascular disease
- Diabetes mellitus
- Alcoholism
- Drug therapy
- Psychogenic impotence
- Sex hormone abnormalities
- Coincident endocrine disease

BOX 7.12 Major sex hormone and prolactin abnormalities in males with CKD

- Low plasma total and free testosterone concentrations
- Low plasma total and free oestrogen concentrations
- Low plasma total and free adrenal androgen concentrations (with normal sex hormone binding globulin concentrations)
- Subnormal plasma testosterone response to human chorionic gonadotrophin administration
- Hyperprolactinaemia in 50% of patients
- Subnormal prolactin responses to TRH in 50% of patients
- Disturbance of the circadian rhythm of prolactin secretion in 50% of patients

and Sertoli cell function. The identity of these 'toxins' remains undetermined, although there is good evidence to suggest that both PTH and prolactin may play a role.

The course of progressive deterioration in testicular function is not markedly altered by either haemodialysis or peritoneal dialysis. The best treatment is renal transplantation, which restores plasma gonadotrophin and sex steroid concentrations to normal and may result in effective spermatogenesis. The immunosuppressive therapy does not appear to have any deleterious effects apart from suppression of adrenal androgen secretion by prednisolone, although there is some evidence that men treated with sirolimus based regimens have a lower sperm count than those who are not.

Other notable problems that occur in dialysed males include gynaecomastia and priapism. The former does not appear to be related to either prolactin concentrations or androgen:oestrogen ratios, and has been compared to the type of gynaecomastia that occurs in malnourished patients during refeeding. The cause of priapism is unknown, but may be related to hypovolaemia after dialysis, the use of heparin or treatment with androgens.

In children, delayed bone age and delayed puberty are found. In boys, Leydig cell function appears to be relatively normal, as shown by the normal testosterone:dihydrotestosterone ratios, suggesting that 5 α -reductase activity is normal. As might be expected, this results in normal LH concentrations, but FSH concentrations are elevated, which appears to reflect damage to the germinal cell epithelium prior to the initiation of spermatogenesis. These observations suggest that Sertoli cell damage in renal impairment occurs early while Leydig cell damage either does not occur until late or only occurs in adult patients.

In women with CKD, there is typically hypothalamic anovulation with amenorrhoea along with loss of libido and inability to reach orgasm. These women frequently exhibit hyperprolactinaemia with reduced oestrogen concentrations and elevated plasma concentrations of gonadotrophins and gonadotrophin releasing hormone (GnRH). Even in women who menstruate, the mid-cycle LH surge appears to be absent leading to anovulatory cycles and low progesterone concentrations, which in turn may cause dysfunctional bleeding. It is not surprising, therefore, that such patients are infertile. With treatment, however, premenopausal women may resume normal menstruation with ovulatory cycles and may even

become pregnant. As with males, renal transplantation produces the most satisfactory results.

Thyroid abnormalities

Patients with CKD have several clinical disturbances that could potentially affect thyroid function. For example, they are chronically ill, often malnourished and display multiple hormonal and metabolic derangements. It is, therefore, not surprising that various abnormalities of thyroid function have been well documented (Box 7.13). These may complicate the diagnosis of coexistent thyroid disease (there is an increased prevalence of goitre and an increased incidence of primary hypothyroidism in CKD, particularly in women). The development of subclinical thyroid function test abnormalities has been explained by a mixture of defects (Box 7.14), which may also be observed in other chronic illnesses and malnutrition states (see Chapter 19). The high incidence of goitre in uraemic patients suggests that there is a circulating goitrogen in CKD. This does not appear to be either TSH or autoimmune antibodies, but there is circumstantial evidence to suggest that PTH may be involved since the incidence of thyroid nodules and goitres is greatly increased at post-mortem in patients with primary hyperparathyroidism.

With peritoneal dialysis (PD), plasma total T4 concentrations are not as low as with haemodialysis (HD), but T3 concentrations are lower. The low T4 concentrations in PD may relate to low concentrations of thyroxine binding globulin (TBG), while in haemodialysis there may be an inhibitor of binding present. Continuous ambulatory peritoneal dialysis (CAPD) appears to correct unbound, plasma free T4 concentrations better than HD; TSH is normal in both groups. Following successful transplantation, the plasma concentrations of thyroid hormones return to normal, while the TSH response to TRH may either be normal or decreased.

BOX 7.13 Plasma concentrations of thyroid hormones and binding globulin in CKD

- Total T4 normal or decreased
- Total T3 decreased
- Free T4 decreased
- Free T3 decreased
- Reverse T3 normal or decreased
- Blunted and/or delayed TSH response to TRH
- Thyroxine binding globulin normal

BOX 7.14 Mechanisms contributing to subclinical thyroid abnormalities in CKD

- Defective thyroid hormone synthesis
- Minor abnormalities in pituitary feedback mechanisms
- Reduced peripheral conversion of T4-T3
- Altered transport of thyroid hormones secondary to circulating inhibitors of binding

The diagnosis of thyroid disease in patients with CKD is difficult. A diagnosis of primary hypothyroidism can only be made with confidence if the plasma TSH is unequivocally elevated. In less clear-cut cases, measurement of free T4 may be helpful, but can be unreliable. The response to treatment should be monitored using plasma free T4 concentrations. The diagnosis of hyperthyroidism in renal impairment rests on the demonstration of an elevated plasma free T4 concentration in the presence of suppressed TSH, although hyperthyroidism is extremely rare in patients with this condition.

Anaemia

In patients with established renal failure, plasma erythropoietin concentrations are usually normal or even elevated. However, these concentrations are generally inappropriately low for the degree of anaemia and so there exists a state of relative erythropoietin deficiency, which contributes to the normochromic normocytic anaemia. Other factors include depression of the erythroid marrow by 'uraemic toxins', bleeding and haemodilution due to water retention. Following renal transplantation, the anaemia improves and the plasma erythropoietin concentrations return towards normal. However, very occasionally they remain elevated and erythrocytosis develops. This is due to some residual secretion of erythropoietin from the diseased kidneys.

Endocrine control of salt and water balance

Chronic kidney disease influences the activity of a number of renal and extrarenal hormones that control salt and water balance as well as blood pressure and volume. Renin release appears to be maintained by diseased kidneys under many circumstances. Thus, depending upon fluid and electrolyte balance, the plasma renin activity (PRA) in patients with renal impairment may be low, normal or high.

Plasma renin activities and 18-hydroxycorticosterone concentrations increase in patients on CAPD more than in those treated with HD. These increased concentrations have been attributed to the continuous fluid volume depletion induced by peritoneal ultrafiltration. Aldosterone concentrations are also elevated secondarily to the hyperkalaemia; they are more effectively reduced by HD than CAPD. Renal transplantation returns PRA and aldosterone towards normal. In addition, the normal physiological responses to sodium restriction, upright position, hypo- or hypervolaemia and angiotensin-converting-enzyme inhibitors appear to be preserved.

Plasma cortisol concentrations are at the upper limit of normal in CKD and respond normally during a tetracosactide test or insulin-induced hypoglycaemia, although the prolactin and GH responses are blunted during the latter, consistent with a degree of anterior pituitary dysfunction.

Plasma catecholamine concentrations are usually elevated in advanced CKD owing to a combination of several mechanisms, including reduced clearance, catabolism and neuronal uptake, and increased sympathetic efflux. After renal transplantation, normal plasma

concentrations of noradrenaline (norepinephrine) with moderately elevated concentrations of adrenaline (epinephrine) have been found. Arginine vasopressin (AVP) is normally freely filtered by glomeruli and catabolized by tubules, so plasma AVP concentrations are elevated in CKD. The hormone is not significantly cleared by dialysis and remains elevated in response to the chronic fluid depletion. Following transplantation, both normal and elevated AVP concentrations have been reported. The plasma concentrations of natriuretic peptides (e.g. BNP, see Chapter 4) rise in direct relationship to the rise in plasma creatinine and represent a normal homeostatic response to fluid overload. Once effective dialysis is established or renal transplantation performed, the concentrations return towards normal.

Impairment of the kidneys' regulatory function results in defects in both the excretory and conservatory mechanisms for sodium and water (Fig. 7.7).

Thirst and nocturia are common features of chronic renal disease, reflecting the diminished capacity to concentrate the urine and loss of the normal diurnal variation in urine production. These effects are primarily due to the solute load causing osmotic diuresis in the remaining functional nephrons. Patients tend not to complain of polyuria; the low GFR sets a low limit on the maximum urine volume of not more than about 4L/24h. The capacity to dilute the urine may be preserved longer than that to concentrate it, but in advanced disease, neither dilution nor concentration can take place. The urine tends to have a fixed osmolality (~300 mmol/kg) and the patient is at great risk of both over- and under-hydration if fluid intake is not carefully regulated.

Early in the course of CKD, normal sodium balance may be maintained by increased excretion through the remaining functional nephrons. Particularly in patients with primarily tubular disorders, sodium conservation may be defective, leading to sodium depletion with clinical evidence of decreased ECF volume. Later, the tendency is to sodium retention, but this may not be apparent until the GFR falls below 10 mL/min and, even then, some patients continue to leak sodium. Since a fall in ECF volume can further impair renal function, careful attention to the maintenance of sodium balance is essential in the management of patients with CKD.

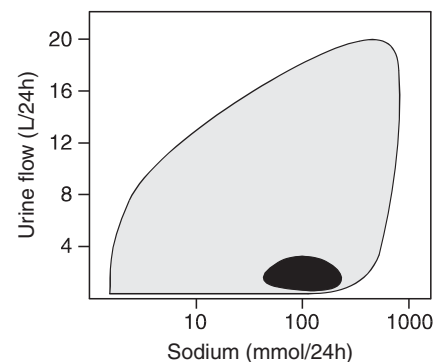


FIGURE 7.7 ■ The renal capability to excrete sodium and water in healthy individuals (light shading) and in patients with severely impaired renal function (dark shading). (From Davison AM et al. (eds) 2005 Oxford Textbook of Clinical Nephrology, 3rd edn. Oxford: Oxford University Press p. 1705, with permission).

Carbohydrate metabolism and lipid metabolism

Impaired glucose tolerance secondary to insulin resistance is a feature of CKD. A number of mechanisms appear to contribute to the generation of this metabolic abnormality (Box 7.15), which is only partially corrected by dialysis and may persist following renal transplantation. However, this may well be related to the effect of steroid therapy since glucocorticoids are well known for their ability to cause hyperglycaemia. Some 70% of patients with CKD exhibit abnormalities of both lipids and lipoproteins, thought to be secondary to the hyperinsulinism. Typically, there is elevation of plasma triglycerides, very low density lipoprotein and intermediate density lipoprotein, but normal total cholesterol, with high density lipoprotein being decreased. The apoprotein and lipid content of the lipoproteins is also altered.

The major defects underlying these abnormalities appear to be low activities of tissue lipoprotein lipase, hepatic lipase and lecithin-cholesterol acyltransferase. These abnormalities undoubtedly contribute to the increase in cardiovascular morbidity and mortality observed in patients with CKD.

Management

General management

This entails measures to evaluate and manage comorbid conditions, slow progression of kidney disease, reduce the risk of cardiovascular disease, prevent and treat complications of chronic kidney disease and to prepare for eventual renal replacement treatment. The measures specific to kidney disease are discussed further below. Management of comorbid conditions will not be discussed further here.

The staging of chronic kidney disease as defined by the chronic kidney disease K/DOQI classification (see Table 7.2) is used as the basis for the development of clinical management plans for individual patients.

Slowing the progression of kidney disease. If the cause of the CKD can be identified, it should of course be treated appropriately. Care must be taken to avoid conditions that may accelerate the decline in renal function, for example dehydration and infection (see Box 7.6).

BOX 7.15 Mechanisms contributing to impaired glucose tolerance in chronic kidney disease

- Metabolic acidosis
- Decreased cellular sensitivity to insulin
- Decreased glucose utilization by peripheral tissues
- Increased plasma concentrations of GH and glucagon
- Retention of Mg^{2+}

Dietary measures are important both in an attempt to retard the progression of disease and in the avoidance of complications. Dietary management should ideally be supervised by a dietitian with special training in the management of renal disease. This person will not only be able to give advice on an appropriate diet, but should be able to design one that is acceptable to the patient and monitor compliance. The essentials of dietary management include maintenance of sodium, potassium and water balance, which usually entails restriction of sodium and potassium intake; reduction in phosphate intake, and dietary protein restriction (to about 0.8 g/kg per 24 h).

Experimental work suggests that judicious reduction in protein intake may decrease the progression of CKD, even when this is well established (plasma creatinine concentration $>400 \mu\text{mol/L}$). Dietary protein restriction may protect against the progression of CKD by haemodynamically mediated reductions in intraglomerular pressure and by changes in cytokine expression and matrix synthesis. The haemodynamic effects of protein-induced hyperfiltration may be due to changes in hormones (such as glucagon and insulin-like growth factor-1), alterations in the renin-angiotensin system, and intrarenal effects, including tubuloglomerular feedback. The benefits of moderate dietary protein restriction (0.6–0.8 g/kg per day) on the progression of CKD in humans remain controversial. Current data suggest that, at best, a small reduction in the rate of decline of glomerular filtration rate (GFR) may be observed with a low protein diet.

When planning nutrition in patients being treated with dialysis, account must be taken of the loss of protein and amino acids into the dialysate. This occurs in patients both on peritoneal dialysis and haemodialysis, although the protein loss tends to be larger with the former and may be up to 40 g/24 h.

Supplementary vitamins may need to be prescribed both for patients treated conservatively and, particularly, those on dialysis. Careful clinical and biochemical monitoring of nutritional status is essential. Adequate energy intake must be maintained (at least 35 kcal/kg per 24 h) to meet energy requirements and prevent catabolism.

The need for careful attention to sodium and potassium intake should be obvious from earlier sections. There are experimental data to suggest that phosphate restriction is also desirable to prevent the development of renal metabolic bone disease, for which purpose the maintenance of plasma phosphate concentration within the reference range is the ideal. Although dietary protein restriction usually also entails dietary phosphate restriction, adequate control may be difficult to achieve. It is also important to avoid decreasing calcium intake, which may contribute to the development of bone disease. Oral aluminium salts can be given to bind phosphate in the gut and reduce its absorption; the risk of aluminium-induced neurological and bone toxicity that led to avoidance of their use may have been overstated, but if they are used, it is important to monitor the serum aluminium concentration regularly. Calcium containing phosphate binders

such as calcium carbonate are widely used, but care must be taken to avoid inducing hypercalcaemia and there is a worry that they may increase the risk of vascular calcification. Newer, non-calcium-based phosphate binders are now available.

Dyslipidaemia is frequently present in patients with CKD. There are conflicting data concerning the effect of lipid lowering therapy with statins on progression of CKD. Some studies suggest that statins slow the rate of decline in renal function in patients with mild-to-moderate renal dysfunction, while others have found that they do not. All the data evaluating the effects of statin therapy on CKD progression were based on subset analyses of trials designed to evaluate the efficacy of statin therapy on cardiovascular disease. Thus, statin therapy cannot be recommended solely for renal protection.

Prevention of complications. It is essential to avoid dehydration. Particular care with fluid balance will be required if a patient has abnormal losses, for example because of vomiting, diarrhoea, or too vigorous treatment with diuretics. Blood pressure must be monitored regularly and any hypertension treated adequately. Any infections, but particularly those of the urinary tract, must be treated promptly, preferably with non-nephrotoxic drugs. Indeed, all medication must be chosen carefully to avoid nephrotoxicity.

The prevention of bone disease involves minimizing hyperphosphataemia (see above) and the administration of 1-hydroxylated metabolites of vitamin D, with or without calcium supplements. If hypercalcaemia occurs owing to tertiary hyperparathyroidism and patients are unfit for surgery, calcimimetics (e.g. cinacalcet) can be used. These drugs act directly on the calcium-sensing receptor in the chief cells of the parathyroid glands to reduce PTH synthesis and secretion.

Recombinant erythropoietin is available for the treatment of anaemia. It is effective and can greatly increase the well-being of patients with CKD, whether treated conservatively or by dialysis. It is, however, expensive and care is needed with patient selection and with monitoring the response to ensure efficient use of the drug. Iron stores must be replete for erythropoietin to work effectively.

Renal replacement treatment

By definition, the patient reaching end-stage renal failure requires renal replacement treatment in order to survive. In the face of such severe loss of renal function, the general measures outlined above will no longer be sufficient to maintain an internal environment compatible with the maintenance of vital functions. There are essentially two options: dialysis (peritoneal or haemodialysis; haemofiltration and haemodiafiltration are related techniques) and renal transplantation. These are not exclusive; most patients will need to undergo dialysis until transplantation becomes possible and, even after transplantation, some will require dialysis until adequate graft function is established or may need it again should the graft fail.

The optimum time to start a patient on dialysis will depend on several factors:

- failure of conservative measures to allow a patient to continue with the activities of daily life, including working
- the onset of major complications (pericarditis, neuropathy, encephalopathy, persistent hyperkalaemia)
- other life-threatening or intolerable features of uraemia.

Early initiation of maintenance dialysis is now preferred to a protracted period of dietary restriction. According to the K/DOQI guidelines (see Table 7.2) dialysis should be initiated electively in those whose eGFR is <15 mL/min. Some patients in the UK are eligible for pre-emptive transplantation if dialysis is predicted to be required within six months (i.e. eGFR is predicted to fall to <15 mL/min).

Haemodialysis. The dialyser consists of two contiguous compartments separated by a semi-permeable membrane. Blood flows through one compartment and dialysis fluid flows in the opposite direction through the other compartment. Waste products diffuse down the concentration gradient from blood to dialysate; water moves according to the relative osmolalities of plasma and dialysate, and other substances that can pass through the membrane (e.g. sodium, potassium) do so in relation to their own gradients (Fig. 7.8). The most frequently employed dialysers consist of an array of hollow fibres with semi-permeable walls; blood is pumped through the fibres, which are surrounded by dialysate.

The composition of the dialysate is similar to that of interstitial fluid, but with a lower potassium concentration and a higher glucose concentration (to make the fluid hypertonic and thus remove water).

Most patients dialyse for 12–15 h/week, spread over three sessions, at home, in a satellite unit or in hospital. It may be possible to relax the dietary restrictions required during conservative management somewhat after starting dialysis, but while more protein (e.g. 1 g/kg per 24 h) may be permitted, potassium, sodium, phosphate and water intake will still require restriction.

Clinical assessment provides an important guide to the adequacy of dialysis, but is inadequate on its own. Measurements of pre- and post-dialysis urea and creatinine concentrations also have limitations. Many alternative means of assessment have been studied, of which ‘urea kinetic modelling’ (UKM) is the most widely used. Measurements of plasma urea concentration are used to calculate the total urea clearance (K) (residual clearance plus clearance by the dialyser, in mL/min). This can be used to calculate the function ‘Kt/V’, where t is the dialysis time in minutes and V the volume of distribution of urea in mL (equal to the total body water or 65% of body weight). The ideal value of Kt/V is >1.0. However, urea is not an ideal marker of renal function; estimates of V may underestimate true values and the desired clearance and actual clearance achieved may not be the same. Furthermore, it is not clear whether increasing benefit may result from the achievement of higher values of Kt/V. Thus, the use of Kt/V to guide dialysis therapy is not as straightforward as it might appear.

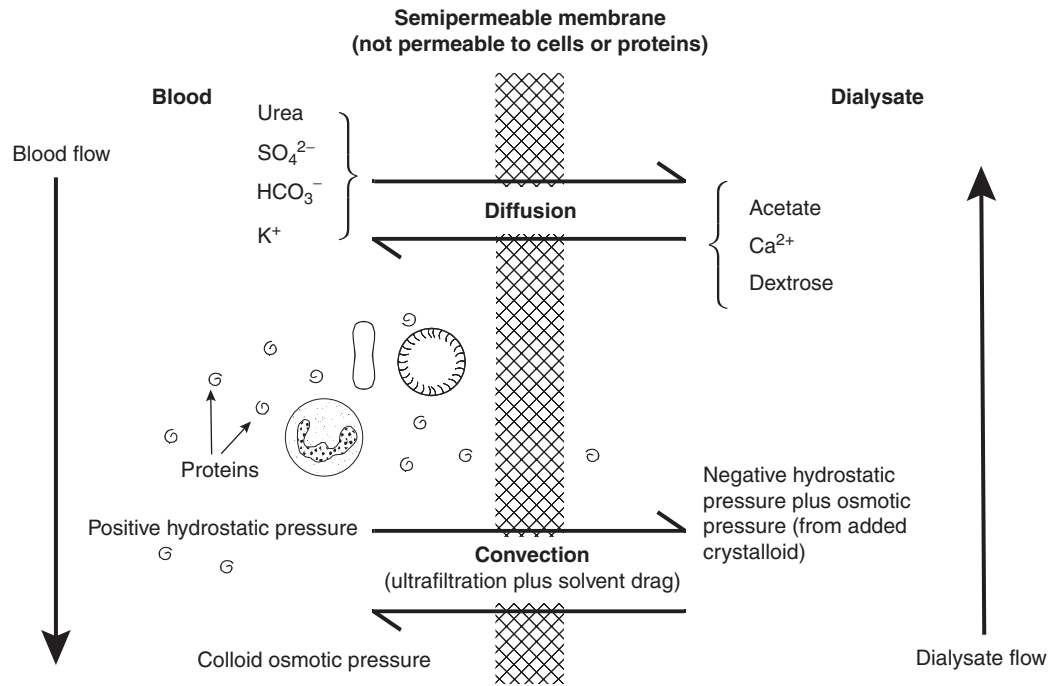


FIGURE 7.8 ■ The principles of haemodialysis.

Haemofiltration. While in haemodialysis, waste products are removed by diffusion, in haemofiltration they are removed by convection. Fluid crosses the filtration membrane as a result of the pressure difference. Haemofiltration membranes are generally more permeable than those used in dialysis. This has the potential advantage that larger molecules (e.g. low molecular weight polypeptides) can be cleared from the plasma than is possible with haemodialysis.

There are systems that incorporate both diffusion and convection, that is, haemodiafiltration. In haemofiltration and haemodiafiltration, the volume ultrafiltered is replaced by intravenous infusion of pre-prepared replacement fluid from bags. Replacement fluid can also be prepared on line by the dialysis machine, for which the water quality has to be ultra-pure.

Peritoneal dialysis. In this technique, the semi-permeable membrane is the peritoneal membrane. Dialysate is instilled into the peritoneal cavity through a permanent silastic dialysis catheter, allowed to equilibrate and then removed. Continuous ambulatory peritoneal dialysis, in which the process is continuous with three or four fluid changes taking place over 24 h, is widely used.

Automated peritoneal dialysis (APD) is a term used to refer to all forms of PD using a machine to assist with the delivery and drainage of the fluid. Exchanges are typically carried out overnight, which leaves the patient free during the day.

Residual renal function can contribute as much as one-third of creatinine clearance; however, it declines with time so the dialysis prescription needs to be kept under review to ensure adequate clearance.

Renal transplantation. In addition to removing the requirement for repeated dialysis, successful transplantation removes the restrictions on the patient's food and fluid intake. The patient will have to take immunosuppressive therapy for life, and, although this poses some risks (increased susceptibility to infection and incidence of lymphoproliferative disorders), they can be minimized by careful monitoring, including therapeutic monitoring of plasma immunosuppressive drug concentrations.

There are many immunosuppressive regimens. Briefly, they include a monoclonal antibody against the interleukin-2 (IL-2)-receptor (e.g. basiliximab or daclizumab) on induction, a calcineurin inhibitor (cyclosporin or tacrolimus), an antiproliferative agent (azathioprine or mycophenolate mofetil) and corticosteroids. Sirolimus is another immunosuppressive drug that is not nephrotoxic; calcineurin inhibitors are. Therapeutic monitoring of these drugs is essential (see Chapter 39). If there is graft dysfunction, toxic concentrations of calcineurin inhibitors are an important reversible cause.

Many factors influence patient survival, but rates continue to improve, and one-year survival rates of >95% for patients are now common. Typically, >88% of deceased donor grafts and >95% of live related grafts are still functioning at one year.

Following transplantation, careful monitoring is required to assess graft function, monitor the patient's hydration and to assist the early diagnosis of rejection. Regular measurements are required of urine volume, plasma creatinine, sodium, potassium, bicarbonate, glucose, calcium, phosphate and albumin concentrations, liver function tests, haemoglobin and white blood cell and platelet counts, as well as therapeutic monitoring of immunosuppressive drugs.

CONCLUSION

The kidneys have an essential role in the regulation of the composition and volume of the body fluids and in the excretion of waste products of metabolism. They also have important endocrine functions. They can be affected by disease processes confined to the kidneys, but they are frequently affected by primarily extrarenal disease.

The informed use of biochemical tests plays an important part in the investigation of renal integrity and function, although many other types of investigation are also used in the management of patients with renal disease.

Renal impairment can be acute (when it is usually potentially reversible) or chronic (when it usually progresses inexorably to established (end-stage) renal failure). In both types, there are profound systemic consequences. Conservative measures are important in management, but renal replacement treatment, by some form of dialysis or haemofiltration, is often required. In established (end-stage) renal failure, death will ensue unless the patient either undergoes successful renal transplantation or receives long-term renal replacement treatment.

Note about terminology

The terms 'acute kidney injury' and 'chronic kidney disease' have now replaced 'acute renal failure' and 'chronic renal failure', respectively.

ACKNOWLEDGEMENT

We would like to acknowledge the contribution of Sui Phin Kon and William Marshall, who wrote the chapters for previous editions of the book.

Further reading

Barrett KE, Barman SM, Boitano S et al. *Ganong's review of medical physiology*. 24th ed. New York: Lange Medical Books/McGraw-Hill; 2012.

A concise undergraduate textbook of physiology, which provides a good overview of renal physiology.

Davison AM, Cameron JS, Grünfeld J-P et al. editors. *Oxford textbook of clinical nephrology*. 3rd ed Oxford: Oxford University Press; 2005.

A comprehensive, three-volume work which, while emphasizing the clinical aspects of renal diseases, also includes authoritative accounts of their pathophysiological basis.

The Renal Association. <http://www.renal.org>.

The website of The Renal Association, which has further information about eGFR and chronic kidney disease.

Proteinuria

Anne Dawnay

CHAPTER OUTLINE

INTRODUCTION 152

PROTEIN CONSERVATION BY THE KIDNEYS 152

- The glomerular capillary wall 153
- The theory of molecular sieving 153
- Tubular reabsorption of proteins 154
- Tubular secretion of proteins 154

NORMAL URINARY PROTEIN CONTENT 155

- Determinants of urine protein excretion 155

PROTEINURIA IN KIDNEY DISEASE 156

- Proteinuria in staging and prognosis of chronic kidney disease 156
- Glomerular proteinuria and nephrotic syndrome 156

- Tubular proteinuria 160
- Proteinuria of prerenal origin 162

MICROALBUMINURIA AS A MARKER OF RISK 164

- Microalbuminuria and risk of diabetic complications 164
- Cardiovascular risk 164
- Microalbuminuria as a risk factor in other inflammatory processes 165

CLINICAL INVESTIGATION OF PROTEINURIA 165

- Urine dip-sticks 165
- Collection of urine 165
- Urine protein measurement 165
- Stepwise investigation of proteinuria 166

INTRODUCTION

For over 150 years, proteinuria has been recognized as one of the cardinal signs of renal disease and it continues to be used for the diagnosis and management of patients with nephropathy. However, the use of sensitive immunoassay methods to measure urine albumin has demonstrated pathologically significant proteinuria well below the detection limit of chemical methods. Since the discovery, several decades ago, that low-level albuminuria (microalbuminuria) identifies diabetic patients with incipient nephropathy, further large-scale studies have revealed that microalbuminuria is also a risk factor for macro- and microvascular disease, in both diabetic and non-diabetic populations. Studies over the last ten years have confirmed that microalbuminuria is not only associated with cardiovascular disease but also with any acute inflammatory condition, and is a reflection of systemic vascular endothelial dysfunction. Of particular importance, is the finding that, in some conditions, microalbuminuria is reversible with interventions that protect the vascular system. Thus modern assessment of proteinuria across the full pathological range has a role not just in the diagnosis and management of primary renal disease, but also as a marker of endothelial dysfunction in a variety of non-renal conditions.

PROTEIN CONSERVATION BY THE KIDNEYS

The kidneys receive approximately 25% of the resting cardiac output, which represents approximately 1.2L/min of blood or 650mL/min of plasma. The kidneys' capacity to conserve protein can be judged from a simple calculation. Every 24h, approximately 930L of plasma containing about 70g/L of protein pass through the kidneys, representing 65 kg of protein, of which <100mg (0.00015%) appear in the urine.

The filtration process is dependent on adequate renal blood flow, which is preserved by an autoregulatory system, despite variations in blood pressure (see p. 127). This mechanism allows vasodilatation as perfusion pressure falls and vasoconstriction as pressure rises. The mediators of this process include prostaglandins, kinins and atrial peptides (vasodilators), and angiotensin II, α -adrenergic hormones, thromboxane A₂, noradrenaline (norepinephrine) and vasopressin (vasoconstrictors). In addition, renal arterioles respond within seconds to changes in vessel wall tension; thus, when renal perfusion pressure rises, vessel wall tone increases and, conversely, when renal perfusion pressure falls abruptly, there is a compensatory decrease in vessel wall tone. This phenomenon is called the myogenic reflex and helps to maintain a constant renal blood flow across a range of perfusion pressures.

A minimum intraglomerular pressure, derived from the pumping action of the heart, is required to overcome the two main opposing forces to filtration: the colloidal oncotic pressure and the hydrostatic pressure in the Bowman space. When the renal perfusion pressure falls below 50–60 mmHg, further vasodilatation does not occur and renal blood flow declines in proportion to the reduction in renal perfusion pressure. These mechanisms maintain renal blood flow, and hence glomerular filtration, independently of the normal fluctuations of blood pressure. However, recent studies suggest that chronic hypertension impairs the renal autoregulatory mechanism, which may contribute to hypertension-associated renal damage and proteinuria.

The glomerular capillary wall

The glomerular membrane consists of a modified capillary wall comprising endothelium, an acellular basement membrane and an outer specialized epithelial cell layer (Fig. 8.1). The endothelial cells are thin and fenestrated with 50–100 nm pores. The basement membrane, comprising the lamina rara interna, lamina densa and lamina rara externa, is around 300–350 nm thick and is best considered as a gel-like structure containing 3–5 nm long fibrils but no detectable pores. The numerous foot processes of the epithelial cells (also called podocytes) interdigitate and envelop the outer surface of the glomerular membrane. The foot processes are separated by slit diaphragms about 55 nm in width, which form the final barrier to plasma proteins.

The whole of the glomerular membrane carries a fixed net negative charge, which is partly owing to a glycosialoprotein coat covering both endothelium and epithelium. The charge increases in density from the lamina rara interna towards the lamina rara externa, with the greatest density at the slit diaphragms of the epithelium.

The glomerular capillaries act as high pressure filters, allowing water and low molecular weight solutes to pass

through freely and almost completely retaining plasma proteins. The concentration gradient of proteins across the glomerular membrane produces the colloidal oncotic pressure that must be overcome for filtration to take place.

The theory of molecular sieving

The glomerular membrane selectively allows passage of water and low molecular weight solutes into the tubule and restricts passage of larger molecular weight plasma proteins, based on a combination of molecular size, shape and charge. The pore size of the fenestrae of endothelial cells (50–100 nm) is too great to provide a major restriction to the passage of many proteins, which pass through to the glomerular membrane. Like the glomerular membrane, most plasma proteins carry a net negative charge, resulting in their electrochemical retention by the kidneys. The capillary side of the glomerular basement membrane (lamina rara interna) has a charge density of 35–45 mEq/L, so that electrochemical effects alone reduce the concentration of albumin to 5–10% of that in plasma. Thus, the lamina rara interna of the basement membrane is the first impediment to confront charged molecules, with the lamina densa providing the most effective size barrier to macromolecules. Some macromolecules accumulate at the slit diaphragms of the epithelium where the net negative charge is greatest, and there is evidence that some molecules are pinocytosed by the podocytes.

Based on their molecular radius, shape and charge, different proteins penetrate the glomerular membrane to a variable extent, for example, albumin (radius 3.6 nm, isoelectric point 4.7) is restricted at the lamina rara interna, while lactoperoxidase (radius 3.8 nm, isoelectric point 8.0) may reach the slit pores of the epithelial cells.

Despite the protein-retaining properties of the glomerular membrane, some protein does pass into the

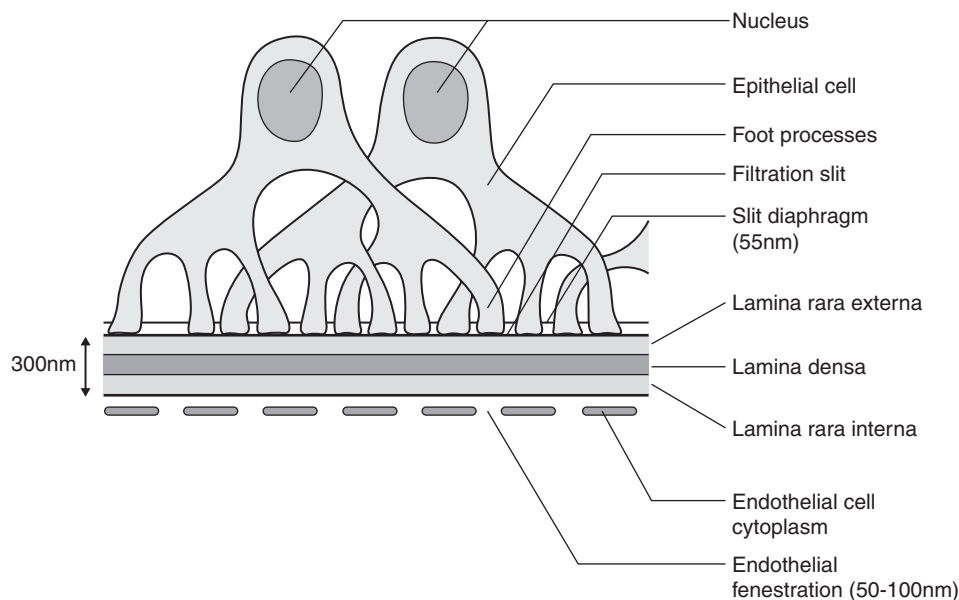


FIGURE 8.1 ■ Glomerular membrane structure.

proximal tubular fluid. In healthy adults, the albumin content of glomerular filtrate is probably about 15 mg/L which, with a glomerular filtration rate (GFR) of 160 L/24h, results in 2–3 g of albumin being presented to the renal tubules each day, most of which is reabsorbed.

Tubular reabsorption of proteins

The amount of protein reaching the proximal tubules is a function of the protein's ability to cross the glomerular membrane and its plasma concentration. After reabsorption, proteins are catabolized in tubular cells and the amino acids, bound vitamins and trace metals returned to the plasma pool. For example, under normal circumstances the kidneys are responsible for about 10% of albumin catabolism, but this figure can rise to 60% when there is increased glomerular clearance of albumin such as occurs in nephrotic syndrome.

Normally, only a very small proportion of plasma proteins reaches the urine. This is primarily owing to retention of large molecular weight proteins by the glomeruli, and almost complete proximal tubular reabsorption and catabolism of any filtered lower molecular weight proteins that have passed through the glomerular membrane more easily. Thus, the renal clearance of plasma proteins represents a combination of these processes. The relative contributions of glomerular retention and tubular reabsorption to protein clearance by the kidney depend on molecular weight (Table 8.1).

The likely mechanism of proximal tubular protein reabsorption is endocytosis at the tubular epithelial apical cell membrane mediated by two receptors (megalin and cubilin) and the cooperating protein amnionless. Mutations in one or other of these components are responsible for the tubular proteinuria of Imerslund–Grasbeck syndrome (selective vitamin B₁₂ malabsorption) and Donnai–Barrow syndrome (a multisystem disorder affecting many organs that normally express megalin). Following reabsorption, proteins are transferred to lysosomes for degradation while the receptors are recycled from endosomes to the apical cell membrane. Mutations in the lysosomal and endosomal components are responsible for the tubular proteinuria in Dent disease and cystinosis; the molecular mechanism in Lowe syndrome is yet to be resolved (see Chapter 9).

Megalin and/or cubilin bind albumin and low molecular weight proteins but there is some evidence that, of the filtered proteins reaching the proximal tubule, those with lower molecular weights are reabsorbed by different mechanisms from those with higher molecular weights.

TABLE 8.1 Relative clearances of plasma proteins

Protein	Molecular weight (Da)	Clearance (%GFR)
IgG	150 000	0.01
Albumin	69 000	0.02
Amylase	48 000	3
Myoglobin	17 800	75
Lysozyme	14 500	80

For example, although increased amounts of low molecular weight proteins are excreted in tubular disease, such large increases are not seen in the urine of patients with massive glomerular proteinuria, when the reabsorptive mechanisms for large molecular weight proteins such as albumin must be saturated. Conversely, experimental induction of massive low molecular weight proteinuria through saturation of reabsorption with other low molecular weight proteins does not induce a similar increase in urine albumin.

Low molecular weight proteins, such as α_1 -microglobulin and retinol-binding protein, that are readily filtered at the glomerulus and normally >99.9% reabsorbed, can be used as sensitive markers of proximal tubular function, since even subtle tubular damage is associated with increased urinary excretion.

Glomerular diseases causing heavy proteinuria induce the release of a multitude of inflammatory and fibrogenic mediators, all of which cause tubulointerstitial fibrosis and renal scarring that contribute to renal impairment. Recent studies suggest that large proteins, such as albumin, which pass through the diseased glomerular membrane, cause tubular cell injury when they reach the proximal tubules. Reabsorption of excessive amounts of these proteins leads to overloading of lysosomes and activation of proximal tubular cells, which produce matrix metalloproteins, cytokines, leukocyte chemoattractants and vasoactive mediators, leading to interstitial inflammation and scarring. The mechanism of injury has been most studied for albumin and it appears likely that compounds attached to the protein, such as fatty acids, are major culprits. In minimal change glomerular disease with heavy proteinuria that is not associated with scarring, the albumin-fatty acid composition appears to be different from that in other glomerulopathies.

Tubular secretion of proteins

Proteins that are too large to be filtered by the normal glomerulus appear in urine either as a result of tubular secretion or from desquamation of tubular epithelial cells as part of the normal cellular turnover. Some large molecular weight proteins may also enter the urine during its postrenal passage along the urinary tract. Measurable activities of enzymes such as N-acetyl β -D-glucosaminidase (150 000 Da), γ -glutamyl transferase (120 000 Da) and lactate dehydrogenase (144 000 Da) can be found in normal urine.

The major urine protein exclusively of renal origin is uromodulin (previously known as Tamm–Horsfall glycoprotein), a heavily glycosylated protein secreted from the thick ascending limb of the loops of Henle. It has a molecular weight of 70 000 Da but is normally present in urine as a large non-covalently linked polymer. Since its isoelectric point is 3.3, it tends to precipitate as a gel during urinary acidification, forming casts by trapping whatever is in the vicinity such as albumin, red blood cells, tubular cells or cellular debris. Numerous mutations in the uromodulin gene have been identified in association with familial juvenile hyperuricaemic nephropathy and medullary cystic kidney disease type 2. These are probably different phenotypes of

the same disease and are known collectively as uromodulin storage disease, owing to accumulation of the misfolded mutated protein in cells of the thick ascending limb of the loops of Henle. They are rare, primary tubulointerstitial disorders with autosomal dominant inheritance characterized by chronic progressive kidney disease, hyperuricaemia and gout, with minimal proteinuria.

The normal biological function of uromodulin remains unclear but it probably has a protective role in trapping potentially damaging material in the urinary space, allowing excretion as casts. Microscopic differentiation of the type of casts present in urine can give useful information about pathological processes within the kidney (see Chapter 7). Experimental studies suggest that uromodulin may have a role in protecting against stone formation. Its capacity to bind type-1 fimbriated *E. coli* has led to the suggestion that its excretion has evolved as a defence against urinary tract infections.

NORMAL URINARY PROTEIN CONTENT

A normal adult excretes about 300 mg/24 h of non-dialysable material, of which <140 mg/24 h is protein, but published reference ranges for total urinary protein excretion vary considerably with the analytical method used. Plasma proteins represent only some 25 mg/24 h of total urinary protein (Table 8.2), of which about half is albumin: the remaining proteins are of renal origin, uromodulin being the major contributor (70 mg/24 h). For these reasons, and because low-level albuminuria (microalbuminuria) has prognostic value both for renal and non-renal diseases, urine albumin immunoassays should be used to assess glomerular proteinuria when protein excretion rates are low.

Determinants of urine protein excretion

Age, sex and diurnal variation

In neonates, albumin excretion tends to be higher than in older children and adults: this has been attributed to greater permeability of the neonatal glomerulus. Total

urine protein tends to fall after birth and rises with increasing age, reaching adult excretion rates by puberty. However, the urine protein/creatinine concentration ratio remains constant from 3 to 15 years of age, since urine creatinine excretion rises with increasing body mass. In children aged 4–16 years, albumin excretion rate, corrected for body surface area, rises with age and is slightly higher in females. Daytime excretion rates are higher than during the night and the sex difference disappears in overnight collections. There appears to be no sex difference in urine albumin excretion in adults, but, when expressed as a ratio to creatinine, the reference range is slightly higher in females owing to their lower muscle mass and hence lower creatinine excretion.

Posture

In both adults and children, ambulatory urine protein excretion is higher than it is overnight or during recumbency, with two- to ten-fold differences being reported for urinary albumin. Orthostatic proteinuria in otherwise healthy subjects has been the subject of controversy for some time. The discussion has been complicated by the variety and differing sensitivities of the protein assays used.

Proteinuria of <1 g/24 h has been described in 0.6–9% of healthy young adults, in the absence of urinary red cells, white cells or casts, and can be divided into ‘constant’ and ‘postural’ based on its persistence after recumbency. Renal biopsies of patients with postural proteinuria reveal that 8% have unequivocal evidence of well-defined disease and 45% have subtle alterations in glomerular structure. However, more recent studies, using non-invasive Doppler ultrasound to compare recumbent and orthostatic blood flow in the left renal veins of young people with orthostatic proteinuria, have revealed that in >50% of patients there is reduced blood flow to the left renal vein during standing owing to entrapment of the left renal vein.

The medical management of isolated or postural proteinuria in an otherwise healthy patient tends to be conservative, with annual assessment of proteinuria and renal function; biopsy is reserved for the rare patient who has evidence of progressive renal impairment.

Exercise and diet

Exercise-induced proteinuria was discovered over a century ago in soldiers after marches or drills. Five- to 100-fold increases in the excretion of proteins such as albumin, transferrin and immunoglobulins have been observed following 26-mile marathon runs, with smaller increases after less strenuous activities. The pattern of exercise-induced proteinuria is generally glomerular, although mixed glomerular and tubular proteinuria has also been described, which persists for over 3 h after exercise. The reason for exercise-induced proteinuria is unclear, but some degree of renal ischaemia owing to redistribution of blood during exercise has been suggested as a possible mechanism.

A large protein meal is associated with an increased urine albumin excretion, which appears to be secondary to an associated increase in GFR.

TABLE 8.2 Plasma proteins found in normal urine

Protein	Mean (range) excretion mg/24h
Prealbumin	0.03 (0.02–0.05)
Albumin	12.7 (8.9–21.9)
α_2 -Acid glycoprotein	0.41 (0.29–0.68)
α_1 -Antitrypsin	0.31 (0.19–0.56)
Caeruloplasmin	0.06 (0.05–0.06)
Haptoglobin	0.18 (up to 0.42)
Transferrin	0.22 (0.13–0.38)
Haemopexin	0.20 (0.14–0.29)
IgA	0.51 (0.37–0.61)
IgG	2.46 (1.97–3.01)
IgM	0.34
Light chains	
lambda (λ)	1.40
kappa (κ)	2.30
β_2 -Microglobulin	0.3

Adapted from Pesce AJ, First MJ 1979 Proteinuria: an integrated review, 1st edn. New York: Marcel Dekker, with permission.

The lowest and most reproducible estimates of urine protein excretion are obtained from an early morning urine specimen after overnight recumbency.

Pregnancy

During normal pregnancy, the urinary albumin excretion rate generally remains within the non-pregnant range, although there is some evidence for a small increase in albumin excretion during the third trimester, which may be related to increased glomerular permeability and/or GFR. Total urine protein excretion increases owing to decreased renal tubular protein reabsorption.

Hypertension in pregnancy is associated with significant maternal and fetal morbidity and mortality. The reliable detection of significant proteinuria is most important in women with new-onset hypertension during pregnancy because it distinguishes between those pregnancies with pre-eclampsia and those with gestational hypertension, the former often requiring admission to hospital owing to the severity of potential complications. In the UK, NICE guidelines for the routine antenatal care of healthy pregnant women recommend blood pressure and urine protein measurement at each antenatal visit; 660 000 women each year will have at least 7–10 such checks.

Gestational hypertension is defined as new hypertension occurring after 20 weeks of pregnancy, but without significant proteinuria. In this group, routine urine protein measurement may be performed using an automated reagent-strip reading device (more reliable than a manual reading) or by a laboratory method. If a reagent strip reading is 1+ or greater, the proteinuria should be quantitated by a laboratory measure in a spot or 24 h urine sample. Women admitted with pre-eclampsia (new hypertension presenting after 20 weeks with significant proteinuria) do not need to have repeated measures of urine protein since there is no strong evidence linking the degree of proteinuria with adverse outcome. Significant proteinuria is defined as >300 mg/24 h or >30 mg/mmol creatinine in a random sample. There are insufficient studies of urine albumin excretion to be able to define cut-offs equivalent to those defining significant proteinuria in gestational hypertension or pre-eclampsia.

PROTEINURIA IN KIDNEY DISEASE

Richard Bright, in 1836, is generally credited with noting the association between proteinuria and kidney disease. Proteinuria remains the most frequent clinical finding and quantitation of proteinuria is valuable for diagnosing, monitoring and assessing the prognosis of kidney disease. Normally, total urine protein excretion is <150 mg/24 h in adults and <140 mg/m²/24 h in children, depending on the methods employed: normal concentrations are often undetectable by chemical methods. However, sensitive immunoassay of specific proteins has extended the detection limits to urine protein concentrations within the reference range. Measurements of low concentrations of specific proteins such as albumin, predominantly reflecting glomerular function, and α_1 -microglobulin or retinol-binding protein, reflecting tubular reabsorptive

function, are now used as very sensitive and early markers of primary renal disease (e.g. glomerulonephritis) and secondary renal disease (e.g. in diabetes mellitus or hypertension). Thus, proteinuria can now be regarded as a continuum that extends from the measurable amount of protein normally excreted in urine up to the 1000-fold increases found in nephrotic syndrome.

Conventionally, proteinuria has been classified into glomerular proteinuria; tubular proteinuria; nephrogenic proteinuria (e.g. uromodulin, basement membrane and tubular proteins); proteinuria of prerenal origin (e.g. overflow proteinurias such as light chain disease, myoglobinuria, haemoglobinuria, lysozyme in leukaemia and amylase in pancreatitis), and postrenal proteinuria owing to obstruction of the urinary tract or inflammation such as occurs in urinary tract infection. This is a convenient way of differentiating the principal sites of the renal abnormality, but is an over-simplification because, for example, glomerular disease leading to large amounts of plasma proteins being presented to the renal tubules leads to inflammatory changes within the tubules and renal scarring. Nor does this classification lend itself easily to the multifactorial causes of proteinuria in some disorders, e.g. HIV-associated glomerular and tubulointerstitial disease that may be modulated by co-infection with other viruses and drug-induced interstitial nephritis.

Proteinuria in staging and prognosis of chronic kidney disease

Numerous epidemiological studies have demonstrated the association of proteinuria with poorer prognosis in people in the general population and across all stages of chronic kidney disease (CKD): any renal disease is more likely to progress, there is an increased risk of developing acute kidney injury, and both all-cause and cardiovascular mortality are increased. These outcomes hold true, whether proteinuria is assessed by dipstick testing or by formal laboratory measurement of either total protein or albumin, in timed urine collections or in random samples. The quantitation of proteinuria (in the absence of a symptomatic urinary tract infection and preferably using the first morning urine) is an essential component of CKD staging, where the suffix 'p' is used to denote its presence. The decision limit is an albumin/creatinine ratio (ACR) >30 mg/mmol (~ 300 mg/24 h) or urine protein/creatinine ratio (PCR) >50 mg/mmol (~ 0.5 g/24 h), although the continuum of risk, albeit lower, extends into the reference range. The presence of proteinuria in CKD is sufficient indication to initiate blockade of the renin-angiotensin-aldosterone system (RAS) with angiotensin-converting-enzyme inhibitors (ACEI) or angiotensin-II receptor blockers (ARB). At higher excretions (ACR >70 mg/mmol or PCR >100 mg/mmol), RAS blockade should be titrated to the highest tolerable levels and referral of the patient to specialist care considered.

Glomerular proteinuria and nephrotic syndrome

In the normal adult, the renal tubules reabsorb about 2–3 g of filtered albumin every 24 h. Thus, even total failure of

this process cannot explain albuminuria of $>3.0\text{ g}/24\text{ h}$; such losses are usually secondary to increased glomerular permeability associated with glomerular damage.

Nephrotic syndrome can be defined as proteinuria severe enough to cause hypoalbuminaemia and oedema. The degree of proteinuria varies but is generally $>3.5\text{ g}/24\text{ h}$ and is accompanied by a plasma albumin $<25\text{ g/L}$. However, it should be remembered that the amount of protein in the urine may decrease as the plasma protein concentration or the GFR falls. Causes of nephrotic syndrome are listed in Table 8.3. In addition to the nephrotic syndrome, glomerular proteinuria is a feature of several other syndromes of nephron injury, and the severity of proteinuria taken together with other clinical findings can allow useful diagnostic classification (Table 8.4).

Mechanisms underlying glomerular proteinuria

Glomerular injury can occur as a result of either primary or secondary kidney disease and there is no single pathogenetic pathway that will embrace all the possible mechanisms. The term glomerulonephritis is generally reserved

for immunologically mediated diseases and excludes other conditions associated with glomerular damage such as diabetes mellitus or amyloidosis.

Glomerulonephritis can be subdivided immunologically into conditions mediated by antibodies against antigens that are either extrinsic or intrinsic to the kidney. Extrinsic antigens include microorganisms causing infections such as bacterial endocarditis or streptococcal infections, and DNA, as in systemic lupus erythematosus (SLE). Antibodies may also develop to an intrinsic structural component, such as the glomerular basement membrane, leading to immune complex formation. Whatever the original mechanism, antigen–antibody complexes become trapped in the glomerulus, activating both the classical and alternative complement pathways and leading to the release of anaphylatoxic components such as C3a and C5a. Anaphylatoxins, together with locally released kinins, prostaglandins and leukotrienes, attract polymorphonuclear neutrophils to the basement membrane, where they release lysosomal enzymes, leading to membrane disruption and glomerular proteinuria.

Examination of renal biopsies, by light and electron microscopy, reveals a gradation of morphological changes with increasing glomerular injury. Loss of anionic charge associated with fusion of the epithelial foot processes can produce a massive but selective proteinuria, while changes in the basement membrane tend to be associated with increasingly non-selective proteinuria (see later).

Minimal change disease. Light microscopy of renal biopsies shows little or no abnormality in this condition, and no immunoglobulin or complement components are seen on immunofluorescence. Electron microscopy shows fusion of epithelial cell foot processes, and the mechanism of proteinuria probably involves loss of the fixed negative charge on the glomerular basement membrane. Although the pathogenesis of minimal change disease has not been completely clarified, clinical and experimental observations suggest that T cell dysfunction plays a role.

TABLE 8.3 Causes of nephrotic syndrome

	Children (%)	Adults (%)
Primary renal disease		
Minimal change (steroid responsive) disease	80	25
Focal segmental glomerulosclerosis	7	9
Membranous nephropathy	1	22
Membranoproliferative glomerulonephritis	10	26
Glomerular disease in systemic conditions		
Amyloidosis	–	7
Diabetic nephropathy	–	3
Systemic lupus erythematosus	–	8
Henoch–Schönlein purpura	2	<1

TABLE 8.4 Summary of clinical correlates of proteinuria

Degree of proteinuria	Clinical correlates
Massive proteinuria	Nephrotic syndrome Proteinuria $>3.5\text{ g}/24\text{ h}$, usually no haematuria, plasma albumin concentration very low, plasma lipids very high, gross oedema, blood pressure normal or low, GFR normal or increased, urine volume normal or increased
Moderate proteinuria	Nephritic syndrome (acute nephritis), acute kidney injury Proteinuria $1\text{--}3.5\text{ g}/24\text{ h}$. Moderate to marked haematuria, plasma albumin normal or low, plasma lipids normal or slightly increased, oedema absent or mild, volume overload sometimes with increased jugular venous pressure, blood pressure raised, GFR reduced, urine volume reduced Chronic kidney disease Similar to nephritic syndrome/acute kidney injury (see above) but over months or years. Moderate proteinuria may reflect reduced renal mass of functioning nephrons rather than extent of renal disease
Asymptomatic, variable/transient	Proteinuria $<1\text{ g}/24\text{ h}$, no other abnormalities, probably benign (orthostatic) proteinuria. Proteinuria $>2\text{ g}/24\text{ h}$, possibly early glomerulonephritis
Persistent low-level proteinuria/microalbuminuria	Associated with early diabetic nephropathy, tubulointerstitial disease, early cardiovascular disease, hypertension, inflammatory conditions, ischaemia and arteriopathy

In the developed world, minimal change disease, frequently of unknown aetiology, accounts for 90% of nephrotic syndrome in 2–6-year-olds and about 20% in adults. In children, there is massive selective proteinuria, which is predominantly albumin with little or no immunoglobulin. This form of nephrotic syndrome usually responds to a short course of prednisolone (steroid-responsive nephrotic syndrome) and a renal biopsy is unnecessary to confirm the diagnosis. The absence of serological evidence for complement activation gives additional confirmation. In adults, the diagnosis is more difficult and a renal biopsy is generally needed before treatment is started. Minimal change disease has a good long-term prognosis with sustained remission and preserved renal function as almost all patients are responsive to treatment.

Membranous nephropathy. This is the most frequent cause of nephrotic syndrome in adults. The common variant is idiopathic membranous nephropathy, in which 80% of patients have plasma antibodies to the phospholipase A2 receptor. These antibodies are rarely found in the secondary forms associated with SLE, hepatitis B virus and cancer. Light microscopy shows thickening of the glomerular basement membrane that is more marked in the later stages of the disease. Immunofluorescence microscopy always shows IgG deposits, but when other immunoglobulin classes (e.g. IgA, IgM) are present, the condition is more likely to be lupus nephritis. Progressive thickening of the basement membrane, demonstrated by electron microscopy, is the most frequent morphological finding in nephrotic syndrome in adults.

Over 80% of patients present with massive proteinuria that is only moderately or poorly selective, while the remaining 20% have asymptomatic proteinuria or microscopic haematuria. The course of the disease is variable, with about one-third of patients going into spontaneous remission and another third progressing to established renal failure. Risk factors for a poor prognosis include severe proteinuria, hypertension, older age, male gender and reduced GFR. Currently, treatment is aimed at lowering blood pressure to <130/80 mmHg with an ACEI or ARB, which usually reduces proteinuria over six months. If proteinuria persists, immunosuppressive treatment is generally started; steroids alone are often ineffective but there remains debate over which additional agents are most effective: these include ciclosporin, chlorambucil and cyclophosphamide.

Membranoproliferative glomerulonephritis. This is characterized by mesangial proliferation and expansion and is not a specific disease but a pattern of injury. As with all glomerulonephritides, many different diseases can produce this pattern of immunologically mediated glomerular damage, including bacterial endocarditis, group A streptococcal infections, hepatitis C virus, monoclonal immunoglobulin deposition disease and SLE. The histological pattern tends to be similar and immunofluorescence microscopy frequently shows IgA, IgG and C3 deposits; circulating complement concentrations are frequently reduced with high concentrations of immune complexes present in some conditions.

Patients with membranoproliferative glomerulonephritis have proteinuria and microscopic haematuria; approximately 50% of those with nephrotic syndrome present with an unselective proteinuria. Management involves identification and treatment of the underlying cause and the use of immunosuppressive agents.

Focal segmental glomerulosclerosis. Hyaline material is deposited in the subendothelial spaces of affected capillary loops of some, but not all, glomeruli, which is why the terms focal and segmental are used. Patients present with mild proteinuria, recurrent haematuria or nephrotic syndrome. The aetiology is not known, although there is some evidence for the presence of a soluble permeability factor and some patients have SLE. Currently, adults are treated with corticosteroids for at least 4–6 months. Efficacy of treatment is judged by monitoring proteinuria; 20–30% will respond with a decrease or cessation of proteinuria. The prognosis in these patients is good, although they may suffer a relapse. Patients whose proteinuria fails to respond to steroids generally go on to develop CKD. Optimal management of this group is currently controversial, but regimens include use of lower-dose corticosteroids with the addition of ciclosporin for 12 months or more.

IgA nephropathy and Henoch–Schönlein purpura. IgA nephropathy is the commonest primary glomerular disease in the developed world; Henoch–Schönlein purpura appears to be a systemic form of the disease occurring primarily in children. A deficiency in glycosylation enzymes leads to the production of galactose-deficient IgA1. This leaves novel glycan residues exposed, to which autoantibodies develop, leading to immune complex deposition. Nephrotic syndrome is uncommon in IgA nephropathy and proteinuria rarely exceeds 5 g/24 h. Treatment involves immunosuppression and antihypertensive therapy with an ACEI or ARB to reduce proteinuria. In Henoch–Schönlein purpura, 50% of children have renal involvement, of whom 20% have nephrotic range proteinuria and associated poorer prognosis.

Urine protein selectivity and classification of glomerulonephritis

The concept of ‘selectivity’ was based on the assumption that there may be differential filtration of large molecular weight proteins among patients with glomerular disease. Protein selectivity is based on a comparison of the relative clearance of IgG (150 000 Da) and transferrin (69 000 Da) calculated as follows:

$$\frac{\text{clearance of IgG}}{\text{clearance of transferrin}} = \frac{[\text{IgG}]_u \times [\text{Trans}]_p}{[\text{IgG}]_p \times [\text{Trans}]_u}$$

where:

- [IgG]_u = urine IgG concentration
- [Trans]_p = plasma transferrin concentration
- [IgG]_p = plasma IgG concentration
- [Trans]_u = urine transferrin concentration.

Where the selectivity index is >0.5 , the proteinuria is said to be 'non-selective', $0.3-0.5$ 'moderately selective' and <0.2 'highly selective'.

Patients with minimal change disease usually show highly selective proteinuria, while patients with membranous nephropathy and membranoproliferative glomerulonephritis do not. Patients with highly selective proteinuria tend to respond well to steroid therapy in contrast to those with non-selective proteinuria, although this is too variable to be useful to determine treatment.

In correlating glomerular disease severity with the type and magnitude of proteinuria, the assumption is made that all nephrons behave in a similar way. However, histological evidence suggests that glomerular disease can be focal with some glomeruli apparently being spared. But if only one nephron out of every 10 000 allowed unrestricted passage of large molecular weight plasma proteins, this would result in an increase in excretion of large molecular weight proteins from 18 to 800 mg/24h, resulting in an unselective proteinuria, apparently indicative of more serious disease.

In addition to glomerular filtration, renal clearances of plasma proteins also reflect tubular reabsorption. If the urine protein content were to represent the overall glomerular permeability, then it must be assumed that tubular reabsorption is entirely unselective. However, small changes in the relative proportion of filtered protein reabsorbed by the renal tubules can have a large effect on their urine concentrations. Despite these limitations, some recent studies have shown that in membranous nephropathy and focal segmental glomerulosclerosis, the selectivity index or IgG excretion predicts remission better than urine total protein. Similarly, progression to CKD is better predicted both by selectivity index or IgG excretion, and the tubular component of proteinuria (e.g. α_1 -microglobulin or retinol-binding protein) than by the total 24h urine protein excretion. However, protein selectivity studies and monitoring of tubular proteinuria are not widely used in the routine diagnosis and management of patients with glomerular disorders.

Pathophysiological consequences of glomerular proteinuria

Hypoalbuminaemia. The severe hypoalbuminaemia found in nephrotic syndrome is not simply secondary to urinary losses exceeding hepatic synthesis, because in adults the maximum hepatic synthetic capacity for albumin is around 14 g/24h, which in theory should replace urinary losses in all but the most severely proteinuric patients. In the nephrotic state, there is a decrease in the circulating albumin pool and hepatic synthesis is generally normal or increased. The loss of the fixed anionic charge on the glomerular membrane, which appears to be fundamental to the glomerular proteinuria of nephrotic syndrome, may also extend to the endothelium of all capillaries, leading to a generalized increase in capillary permeability to albumin.

Oedema and salt and water retention. Until recently, the explanation for the gross interstitial oedema associated

with nephrotic syndrome was attributed solely to hypoalbuminaemia, the consequent reduction in plasma colloidal osmotic pressure being thought to lead to an increase in fluid loss from capillaries at their arteriolar ends and a reduction in tissue fluid reabsorption distally, as predicted from the operation of Starling forces. The resultant hypovolaemia was thought to stimulate the renin-aldosterone system and vasopressin release, leading to sodium and water retention. However, recent work suggests the explanation is more complex.

Normally, about 60% of total body albumin is in the interstitial space and 40% in the vascular space. Every hour, about 8 g of albumin cycle between these two fluid compartments. The gradual loss of plasma albumin from the circulation in nephrotic syndrome leads to a fall in plasma oncotic pressure (the osmotic pressure due to proteins), which in turn is accompanied by a decrease in tissue albumin and parallel fall in interstitial oncotic pressure. The balancing of falling oncotic pressure on either side of the capillary membrane protects against a dramatic fall in blood volume as a result of movement of fluid from the hypo-oncotic vascular space into the interstitium. However, the capacity of this protective mechanism is limited, and a decrease in blood volume can be expected if plasma oncotic pressure (normally 22–28 mmHg) falls below about 10 mmHg, or if the protein loss is very sudden. This is probably because the interstitial oncotic pressure is also maintained by the glycosaminoglycans matrix.

Thus, patients with nephrotic syndrome are not usually hypovolaemic and renal perfusion is generally normal. Despite this, there is intense renal sodium and water retention, which is not mediated by the renin-aldosterone system, and which exacerbates the interstitial oedema. The mechanism is not well understood, but may involve increased sodium reabsorption by epithelial sodium channels in the collecting ducts and decreased sensitivity to natriuretic peptides. In most patients, the oedema is managed by treatment with oral loop diuretics, most often furosemide. If this fails, treatment includes salt restriction, intravenous furosemide, additional thiazide or potassium-sparing diuretics and, occasionally, in severe cases, intravenous albumin.

Abnormalities of other plasma proteins. Patients with nephrotic syndrome have changes in the concentrations of circulating clotting factors, which may be responsible for an increased risk of thrombus formation including renal vein thrombosis, inferior vena cava thrombosis and pulmonary emboli. Urinary loss of the natural anticoagulant antithrombin III is probably involved, together with increased plasma concentrations of fibrinogen and factors V, VII, VIII and X, the reasons for which are unclear.

Urinary loss of vitamin D and vitamin D-binding globulin may lead to vitamin D deficiency with low plasma ionized calcium, secondary hyperparathyroidism and renal osteodystrophy. Some national guidelines recommend vitamin D supplementation, especially in children with nephrotic syndrome. Hypochromic microcytic anaemia, resistant to iron supplements, can occur as a result of urinary loss of transferrin. Patients

with nephrotic syndrome may also lose immunoglobulins and complement components, with attendant increased risk of infections such as pneumococcal pneumonia and peritonitis.

Hyperlipidaemia. Many patients with nephrotic syndrome have massive increases in plasma cholesterol and triglyceride concentrations, which correlate inversely with the plasma concentration of albumin. Plasma very low density lipoprotein (VLDL) and low density lipoprotein (LDL) concentrations are increased, partly owing to reduced clearance and partly to increased hepatic synthesis, while high density lipoprotein (HDL) is reduced because of increased urinary loss. Very low density lipoprotein and LDL are apoB-containing lipoproteins, which are potentially atherogenic and have been linked to the pathogenetic processes that result in progressive glomerular and interstitial renal disease and increased cardiovascular risk. The activities of lecithin-cholesterol acyltransferase (LCAT) and the lipoprotein lipases are reduced in patients with nephrotic syndrome, yet increased urinary losses of these enzymes have not been demonstrated and the underlying mechanism for their reduced activity is not clear. Reduced LCAT activity may also reduce HDL production and hence, impair the patient's ability to mobilize lipid from endothelium and peripheral tissues.

Hepatic synthesis of VLDL and LDL seems to be stimulated by the reduction in plasma oncotic pressure, since treatment with albumin or dextran infusions reduces hepatic lipoprotein synthesis. These abnormalities are summarized in Table 8.5. Nephrotic hyperlipidaemia is accompanied by an increased risk of cardiovascular complications and is usually treated aggressively with statins. The impact of this intervention on cardiovascular morbidity and mortality in patients with nephrotic syndrome has still to be demonstrated.

Tubular proteinuria

Unlike glomerular proteinuria, where protein excretion can reach 20 g/24h and consists mainly of albumin, tubular proteinuria is generally <1–2 g/24h. In tubular proteinuria, although albumin remains a significant component, there is a relatively much greater increase in proteins of <60 000 Da (Table 8.6).

Low molecular weight proteins are filtered rapidly by the glomerulus, and proteins of <15 000 Da enter the Bowman space almost as easily as water and dissolved minerals. Under normal circumstances, the majority of filtered proteins and other solutes are reabsorbed by the proximal tubule, but in renal tubular disease, the reabsorption of protein together with water, ions, glucose and amino acids is impaired. The changing proportions of low and high molecular weight proteins in glomerular and tubular disease, shown in Table 8.6, can be explained by considering the renal handling of albumin (69 000 Da) and β_2 -microglobulin (11 815 Da).

Despite the large differences in plasma concentration of albumin (40 000 mg/L) and β_2 -microglobulin (2 mg/L), about 15 and 2 mg/L, respectively, appear in glomerular filtrate owing to differences in glomerular sieving. Assuming 160 L of filtrate are produced each day, 2400 mg of albumin and 320 mg of β_2 -microglobulin are filtered daily and would be excreted in the absence of tubular reabsorption. The maximum normal daily excretion of albumin is about 20 mg and of β_2 -microglobulin about 0.3 mg. Thus, in tubular disease, the increase in excretion of low molecular weight proteins is far greater than that found for larger proteins such as albumin.

Renal disorders associated with tubular proteinuria

In primary tubular proteinuria and interstitial nephritis, the renal interstitium and tubules are principally affected, rather than the glomeruli or renal blood vessels.

TABLE 8.5 Summary of lipoprotein and apolipoprotein abnormalities in nephrotic syndrome

Lipoprotein fraction	Plasma concentration	Cholesterol content	Triglyceride content
VLDL	↑ ^a	↑	↑
LDL	↑ ^a	↑	Normal
HDL	Normal/↓ ^b	↓	Normal
Apolipoprotein		Cholesterol content and triglyceride content do not apply to apolipoproteins	
AI	Normal/↓ ^b		
AII	Normal/↓ ^b		
Total A	Normal/↓ ^b		
B-48 and B-100	↑		

^aIncreased in more severe nephrotic states: inversely related to plasma albumin concentration.

^bDepends on severity of nephrotic syndrome: lower concentrations occur in severe cases.

TABLE 8.6 Typical urinary protein excretions in glomerular and tubular disease (see text for details)

Protein	Normal	Glomerular disease	Tubular disease
Total protein (mg/mmol)	<15	>250	<200
Albumin (mg/mmol)	<3	>50	<50
α_1 -Microglobulin (mg/mmol)	<1	Normal/slightly ↑	↑↑
β_2 -Microglobulin (μ g/mmol)	<30	Normal/slightly ↑	↑↑

However, recent studies have demonstrated that primary glomerular proteinuria leads to large amounts of plasma proteins being presented to the proximal renal tubules, absorption of which causes inflammation and secondary tubular damage. Thus, in addition to direct tubular toxic effects of conditions mediated by autoimmune or infective mechanisms, drugs, toxins, metabolic disorders or heavy metals, any pathology leading to glomerular proteinuria will also be associated with some degree of tubular dysfunction (Box 8.1).

Measurement of urine total protein or albumin is of limited value in differentiating between glomerular and tubular proteinuria, other than that the latter tends to be associated with proteinuria of <1–2 g/24h. Estimation of proteins that more specifically reflect tubular reabsorption is of more help in identifying primarily tubular disorders. In addition to tubular proteinuria, acute and chronic tubulointerstitial disease can lead to other tubular dysfunction with loss of water, sodium, potassium, calcium, glucose, phosphate, urate and amino acids. Bicarbonate wasting can lead to renal tubular acidosis (type 2). Tubular proteinuria, together with defective reabsorption

BOX 8.1 Causes of tubulointerstitial nephritis

- Drugs
 - Analgesics
 - Non-steroidal anti-inflammatory drugs
 - Cytotoxic agents, e.g. cis-platinum
 - Ciclosporin
- Heavy metals
 - Lead
 - Cadmium
 - Mercury
 - Bismuth
 - Copper (Wilson disease)
- Urinary tract obstruction^a
- Metabolic disease
 - Vitamin D deficiency
 - Hyperuricaemia
 - Hypercalcaemia
 - Tyrosinaemia
 - Hereditary fructose intolerance
 - Galactosaemia
 - Cystinosis
 - Sickle cell anaemia^a
 - Renal tubular acidosis
 - Lowe syndrome (oculocerebrorenal syndrome)
- Infection
 - Urinary tract infection with anatomical abnormality
 - Systemic infections (streptococcal infections, leptospirosis, toxoplasmosis, infectious mononucleosis, brucellosis, syphilis)
- Neoplastic disease
 - Multiple myeloma^a
 - Light chain disease
 - Leukaemias and lymphomas
- Immunological diseases
 - Systemic lupus erythematosus^a
 - Sjögren syndrome^a
 - Transplant rejection^a
 - Amyloidosis^a

^aMay have coexisting glomerulonephritis with proteinuria >1–2 g/24h.

described above, may form part of the Fanconi syndrome, which can be inherited or acquired (see Chapter 9).

Drug and heavy metal induced tubular damage.

In general, the renal tubules are particularly susceptible to drug induced damage because the renal concentrating mechanism leads to high drug concentrations within the tubules. Aminoglycoside antibiotics bind to phospholipids and can disrupt cellular membranes. Analgesic drugs such as phenacetin or its metabolite acetaminophen (paracetamol) damage tubular cells because they are concentrated within the renal medulla, bind to cellular proteins and deplete stores of glutathione. Non-steroidal anti-inflammatory drugs (NSAIDs) inhibit the synthesis of renal vasodilatory prostaglandins, which can lead to unopposed vasoconstriction and ischaemia.

Arsenic, chromium, cadmium, copper, bismuth, mercury and lead have all been described as causing renal injury. Heavy metals react with membrane-bound sulphhydryl groups, altering membrane permeability and causing cellular damage. Inhalation of mercury vapour appears to cause glomerular proteinuria, while ingestion of mercury salts can produce combined glomerular and tubular proteinuria. Acute lead intoxication in children can cause tubular proteinuria, although patients with clinically apparent lead poisoning sometimes have normal protein excretion. Cadmium is taken up by renal tubules as a metallothionein complex that is more toxic than cadmium alone. Increased excretion of β_2 -microglobulin was first described in cadmium workers, and markers of tubular proteinuria are now used to monitor occupationally exposed subjects.

Methods of assessing tubular damage

Renal tubular function can be assessed in a number of ways, such as by tests of reabsorption of glucose, phosphate, bicarbonate and amino acids. Discussion here will be limited to the use of urine proteins as markers of tubular damage by measurement of either high molecular weight proteins, such as enzymes or brush border antigens released by the damaged tubules, or of low molecular weight proteins filtered at the glomerulus and normally reabsorbed by the proximal renal tubule.

High molecular weight protein markers of renal tubular damage.

Several enzymes not normally filtered by the glomeruli are released by damaged tubular cells and can be detected in the urine. They include lactate dehydrogenase, γ -glutamyl transferase and alkaline phosphatase. Their use as markers of tubular damage has been limited because of their instability in urine and the presence of urinary enzyme inhibitors.

The enzyme most widely employed for monitoring tubular damage is N-acetyl β -D-glucosaminidase (NAG), a lysosomal enzyme of 150 000 Da, which is found in high concentrations in the cells of the proximal tubule. Two isoenzymes of NAG are found in urine, one acidic and one basic: the acidic form is found in normal urine and both are excreted in patients with renal disease. Its lack of specificity limits its utility but NAG may be of value in monitoring inherited tubulopathies and in assessing tubular damage during cancer chemotherapy or treatment

BOX 8.2 Conditions associated with increased urinary N-Acetyl β -D-glucosaminidase excretion

- Nephrotic syndrome
- Glomerulonephritis
- Tubulointerstitial nephritis
- Heavy metal poisoning
- Renal transplant rejection
- Diabetes mellitus (with and without nephropathy)
- Hypertensive renal disease
- Surgery
- Ischaemia
- Hyperthyroidism
- Cancer chemotherapy

with antiretrovirals. Conditions associated with increased NAG excretion are given in [Box 8.2](#).

Low molecular weight protein markers of renal tubular disease. Ideally, a low molecular weight urine protein that can be used to monitor tubular damage should have the following characteristics:

- it should have a constant plasma concentration
- be freely filtered by the glomerulus
- its tubular reabsorption should be near saturation so that small reductions in tubular function result in significant urinary excretion
- its tubular reabsorption should be unaffected by co-existing glomerular proteinuria
- it should be stable in urine
- it should be easily measurable.

Since the plasma concentration of low molecular weight proteins filtered by the glomeruli rises as the GFR falls, the tubules of those nephrons still functioning are presented with an increased low molecular weight protein load that causes overflow into the urine and confounds the interpretation of results. The characteristics of some proteins used as tubular markers are given in [Table 8.7](#).

Lysozyme (muramidase) was the first low molecular weight marker of tubular function. Urine lysozyme excretion is increased in patients with Fanconi syndrome, but is often within normal limits in patients with nephrotic syndrome. Plasma lysozyme concentration is increased in

TABLE 8.7 Proteins used as markers of tubular damage

Protein	Molecular weight (Da)	Mean normal urinary excretion ($\mu\text{g}/\text{mmol creatinine}$)	Isoelectric point
Retinol-binding protein	21 000	8	4.7
α_1 -Microglobulin	31 000	400	3.6
β_2 -Microglobulin ^a	11 815	10	5.8

^aNote: unstable at urine pH <5.5. Requires in vivo urine alkalinization to avoid erroneous results.

inflammatory conditions and in some leukaemias, which limits its value as a tubular marker.

Increased excretion of β_2 -microglobulin was first identified in patients with cadmium poisoning and with Wilson disease. This protein is associated with the heavy chain of human leukocyte antigen surface protein on cells and is released during normal cellular turnover. Increased plasma concentrations are found in patients with liver disease and some malignancies such as myeloma and B cell lymphomas. However, β_2 -microglobulin is unstable in acidic urine (pH <5.5) owing to degradation by neutrophil elastase. This necessitates strict control of urine pH after collection, although little can be done to prevent in vivo degradation in the bladder before voiding other than oral administration of bicarbonate to alkalinize the urine.

Retinol-binding protein (21 000 Da) forms an 85 000 Da complex with vitamin A and prealbumin in plasma and hence only a small proportion is free to be filtered by the glomeruli. Plasma concentration falls in inflammatory disease and in vitamin A deficiency. Plasma α_1 -microglobulin (31 000 Da) concentrations are increased in neoplastic diseases, but are reduced in liver disease in parallel with plasma albumin. Unlike β_2 -microglobulin, retinol-binding protein and α_1 -microglobulin are relatively stable in urine and are the most commonly measured markers of tubular proteinuria.

Proteinuria of prerenal origin

Proteinuria of prerenal origin has been defined as the occurrence in the urine of abnormal amounts of protein filtered by the glomeruli in the absence of any glomerular or tubular abnormality. The term is usually applied to 'overflow' proteinuria such as Bence Jones proteinuria, haemoglobinuria and myoglobinuria, where plasma concentrations of these proteins are increased. This definition has limitations because the very presence of abnormally high concentrations of protein at the glomerulus and in the tubular lumen can cause glomerular and tubular abnormalities. Furthermore, it is becoming apparent that non-renal conditions are associated with proteinuria that disappears when the condition resolves.

Myoglobinuria and haemoglobinuria

Myoglobin is found predominantly in skeletal and cardiac muscle cells. Any disease that causes rapid destruction of striated muscle (rhabdomyolysis) results in the release of myoglobin and other muscle proteins into the circulation. Having a molecular weight of 17 000 Da, myoglobin is filtered rapidly by the glomerulus; it has a renal threshold of about 15 mg/L.

[Box 8.3](#) illustrates the diversity of causes of rhabdomyolysis, although a common factor appears to be the damaging effects on muscle cells of a failure to meet increased energy demands. Rhabdomyolysis is associated with acute kidney injury, but a causal relationship is not clear. Although myoglobin casts are found in rhabdomyolysis, infusion of pure myoglobin does not in general cause kidney damage, suggesting that there are other factors that

BOX 8.3 Causes of rhabdomyolysis and myoglobinuria

- Severe exercise
- Injury
 - Trauma
 - Electric shock
 - Crush injury
 - Surgery
- Ischaemia
- Metabolic
 - Severe hypokalaemia
 - Severe hypophosphataemia
 - Glycogen phosphorylase deficiency (McArdle disease)
 - Phosphofructokinase deficiency
 - Carnitine palmitoyl acyltransferase deficiency
 - Malignant hyperpyrexia
- Infections
 - Viral (influenza, coxsackie)
 - Bacterial (typhoid, *Shigella* spp., haemolytic streptococcus)
- Toxins
 - Carbon monoxide (leading to hypoxia)
 - Snake venoms
 - Drugs (steroids, statins, barbiturates, alcohol)
- Dermatomyositis and polymyositis

lead to the kidney injury. Purine derivatives released from damaged muscle cells are rapidly converted to uric acid, leading to very high plasma urate concentrations and intrarenal deposition. Acidosis as a result of the initiating cause of rhabdomyolysis, for example hypovolaemic shock, or from the release of organic acids from muscle cells, may potentiate the degradation of myoglobin to globin and ferrihaemate within the tubules. Ferrihaemate is toxic to tubular epithelial cells. Desquamation adds to the tubular obstruction.

Intravascular haemolysis releases free haemoglobin, which binds to circulating proteins such as haptoglobin. Since haemoglobin itself is 68 000 Da and its plasma protein-complexed form is even larger, haemoglobinuria does not occur until haemolysis is severe and plasma haemoglobin concentration exceeds 1 g/L. Pure haemoglobin is not nephrotoxic, and acidosis and dehydration are important factors determining the nephrotoxicity of severe intravascular haemolytic episodes. Haemoglobin casts are found, and ferrihaemate released from degraded

TABLE 8.8 Biochemical features associated with myoglobinuria and haemoglobinuria

Plasma	Myoglobinuria	Haemoglobinuria
Creatine kinase	↑↑↑	Normal
Visible haemolysis	None	Present
Calcium	↓	Normal
Haptoglobin	Normal	↓
Urate	↑↑↑	Normal
Creatinine	↑ ^a	Normal

^aPlasma creatinine concentration disproportionately increased compared to GFR.

haemoglobin may also cause tubular epithelial cell damage.

Both myoglobin and haemoglobin possess peroxidase activity, which gives positive results for blood with reagent sticks. Measurement of urine myoglobin and haemoglobin is rarely necessary because other biochemical features give better clues to the diagnosis (Table 8.8).

Paraproteinaemias and Bence Jones proteinuria

Bence Jones proteinuria, the presence in the urine of immunoglobulin light chains, is a frequent finding in patients with multiple myeloma and may occur with other B cell malignancies (Table 8.9). This topic is considered in detail in Chapter 30.

There is an association between the presence of Bence Jones protein and the development of renal dysfunction in patients with multiple myeloma. Free light chains (molecular weight 22–24 000 Da) are normally cleared from the circulation by glomerular filtration followed by reabsorption and catabolism in the proximal tubules. Given the normal production rate of 500 mg/day and a reabsorptive capacity of 10–30 g/day, the production rate must increase substantially to cause overflow into the urine in the absence of renal dysfunction. Some light chains are toxic to the proximal tubules (a property that appears to relate to the amino acid sequence in the variable domain but is incompletely understood) and probably contribute to the development of Fanconi syndrome, renal tubular acidosis and renal

TABLE 8.9 Relative frequencies of light chain proteinuria in monoclonal paraproteinaemias

Type	Relative frequency (%)	Light chain proteinuria (%)	Renal impairment (%)
Multiple myeloma			
IgG	55	60–65	50
IgA	20–25	70	60
IgD	1–4	90–100	90
IgE	<1	100	Rare
Waldenstrom macroglobulinaemia			
IgM	14	90	Rare
Light chain disease	10–20	100	90
Heavy chain disease	<1	–	Rare
Benign monoclonal paraproteinaemia	–	<5	0

Adapted from Sweny P, Farrington K, Moorhead J F 1989 The kidney and its disorders. Oxford: Blackwell Scientific, with permission.

failure in patients with multiple myeloma. Patients with light chain proteinuria show increased excretion of lysozyme, retinol-binding protein, β_2 -microglobulin and albumin (with β_2 -microglobulin excretion being higher in λ light chain proteinuria), suggesting that these proteins compete for common reabsorptive mechanisms. Light chains can complex with uromodulin, producing typical myeloma casts, especially when patients are dehydrated and/or acidotic. Partial degradation of light chains can cause polymerization and the formation of damaging amyloid deposits in the glomerular and tubular basement membranes. The specific detection and quantitation of urine paraproteins and Bence Jones protein requires electrophoresis, immunofixation and urine total protein analysis although specific immunoassays for urine free light chains are now available.

MICROALBUMINURIA AS A MARKER OF RISK

The term 'microalbuminuria' was coined to describe an increase in urine albumin that is detectable by sensitive immunoassays but is below the detection limit of chemical urine protein methods and dye-binding stick tests. Microalbuminuria is defined as an albumin excretion rate of 20–200 $\mu\text{g}/\text{min}$ (30–300 $\text{mg}/24\text{h}$ or 3–30 mg/mmol creatinine). Diabetic patients with microalbuminuria are at increased risk of developing dipstick positive proteinuria and CKD. Early identification allows aggressive treatment to improve glycaemic control, blood pressure and plasma lipids and has been shown to improve outcome. More recently, large scale studies have shown that screening for microalbuminuria in non-diabetic as well as diabetic populations identifies patients at increased cardiovascular risk. Furthermore, microalbuminuria has been found to be a predictor of outcome in critically ill patients following insults such as major surgery, trauma or sepsis. The link between microalbuminuria and outcome in such apparently diverse groups of patients appears to be that microalbuminuria reflects the systemic microvascular endothelial dysfunction that, in various forms, is common to all these conditions.

Microalbuminuria and risk of diabetic complications

Diabetes mellitus is the commonest cause of established renal failure. It can affect the kidneys in a number of ways leading to glomerular disease, obstruction, tubular disease and a predisposition to infection, but it is diabetic glomerular disease that is the most important, involving about one third of patients. In the early stages of diabetic renal disease, there is hyperfiltration with an increased GFR, which may be associated with microalbuminuria. With adequate treatment, the GFR falls towards normal and microalbuminuria disappears. However, microalbuminuria returns during periods of poor glycaemic control and becomes persistent with established disease, usually after 5–20 years of diabetes. The risk of developing diabetic nephropathy can be reduced by improving glycaemic control and by treating hypertension aggressively. Once albuminuria reaches

>0.5 $\text{g}/24\text{h}$ or more, improved glycaemic control appears to have little effect on further progression.

Type 1 diabetes usually presents acutely and microalbuminuria rarely occurs within five years of diagnosis, so annual monitoring of urine albumin is generally recommended only after five years' disease duration. In contrast, type 2 diabetes will often have been present for many years prior to diagnosis, hence annual monitoring of urine albumin is recommended from the time of diagnosis. To reduce micro- and macrovascular risk, patients with persistent microalbuminuria (i.e. confirmed by two repeat measurements over 3–4 months) should be treated with an ACEI (or ARB), titrated to the maximum tolerable dose, to reduce intraglomerular pressure. Blood pressure should be maintained at <130/80 mmHg by the use of additional antihypertensives if necessary. These measures, together with aggressive management of cardiovascular risk factors, aim to reduce the rate of progression of kidney disease and prevent cardiovascular events. However, it should be noted that a diminishing urine albumin excretion rate associated with treatment does not invariably translate into prevention of progression of renal disease as assessed by GFR.

Cardiovascular risk

Microalbuminuria is a marker for generalized vascular endothelial dysfunction. In the USA and the European Community, microalbuminuria is found in 6–10% of the general population. Numerous clinical studies in non-diabetic populations have found an association between microalbuminuria and cardiovascular risk factors, target organ damage and the presence of cardiovascular disease. Although microalbuminuria interacts with other cardiovascular risk factors, it is an independent predictor of cardiovascular disease. In large population studies, the presence of microalbuminuria has been shown to be associated with raised plasma C-reactive protein (CRP) concentrations (measured using high sensitivity assays), which is consistent with the view that microalbuminuria reflects systemic endothelial dysfunction mediated by inflammatory processes. Large cross-sectional prospective studies have shown an association between urine albumin excretion and both blood pressure and cardiovascular risk that extends from within the urine albumin reference range through microalbuminuria and to clinical proteinuria. In patients with primary hypertension, the prevalence of microalbuminuria is typically 4–6%. In subjects with mild hypertension and no cardiovascular complications, urine albumin excretion is determined by the haemodynamic load, whereas in subjects with more severe hypertension and associated target organ damage, the urinary albumin leak is probably the consequence of glomerular damage.

Treatment with an ACEI or ARB reduces urine albumin excretion and may prove to be a more targeted approach to reducing cardiovascular risk. Microalbuminuria can be considered not only as a risk factor for progressive renal damage, but as providing an integrated assessment of long-term damage to the cardiovascular system, and is therefore increasingly being used in cardiovascular risk assessment clinics.

What remains to be shown is whether targeted treatment of microalbuminuria in the non-diabetic population reduces cardiovascular morbidity and mortality. However, there is general consensus that detection and quantitation of microalbuminuria is useful for the assessment of overall cardiovascular risk in hypertension, since it appears to be a cost-effective way of identifying patients at higher risk for whom additional preventive and therapeutic measures are advisable.

Microalbuminuria as a risk factor in other inflammatory processes

The emerging association between inflammation, vascular endothelial dysfunction and microalbuminuria has led to a large number of studies exploring the potential clinical value of this relationship. Interaction of activated leukocytes and the vascular endothelium is at the heart of the inflammatory process, and is normally under tight homeostatic control, preventing the local inflammatory response from becoming a systemic, life-threatening process. Yet this homeostasis can fail in patients with severe injuries, infection or multiple pathologies, resulting in systemic inflammatory response syndrome (SIRS). This is the underlying pathogenetic mechanism for the development of multiple organ failure, which is the major cause of death in the critically ill (see Chapter 20). Attempts to modulate the severe inflammatory response have been unsuccessful, partly because of the multiplicity of inflammatory pathways, such that blocking a single pathway is unlikely to be effective, and because such interventions must be given early in the evolution of SIRS to have any hope of success. Increased urine albumin excretion occurs within minutes of any acute inflammatory stimulus such as trauma, surgery, ischaemia reperfusion injury or infection, and normally returns to baseline within 6–8 h, depending on the nature and magnitude of the inflammatory insult. However, in patients who develop SIRS, microalbuminuria is sustained, reflecting systemic vascular endothelial dysfunction. Thus, serial monitoring of urine albumin during major inflammatory episodes has the potential to identify, within a few hours of onset, patients who require close monitoring in an intensive care unit, and who might be candidates for immune-modulating therapy. On the other hand, a return of urine albumin excretion to normal provides a marker of restoration of homeostatic control of inflammation, which can be used as an indicator of successful recovery from the acute inflammatory insult.

CLINICAL INVESTIGATION OF PROTEINURIA

Urine dip-sticks

For many years, proteinuria has been identified by screening random urine specimens using a semiquantitative chemical stick test. These methods are based on the colour change of indicators such as bromophenol blue, which is buffered to pH 3.0 with citrate, and changes from yellow to blue when bound to albumin. Such stick

tests for protein (that predominantly detect albumin) have a detection limit in the range 200–250 mg/L and are subject to errors, including false positives owing to alkaline pH from infection with urea-splitting organisms or contamination with antiseptics, and false negatives owing to very dilute urine and visual reading errors. With the growing awareness in diabetic medicine of the pathological significance of proteinuria at concentrations that are undetectable by conventional chemical sticks, specific semiquantitative immunoassay sticks for urine albumin have been developed that can detect albuminuria at ~30 mg/L.

Collection of urine

Measurement of urine protein or albumin as a concentration in milligrams or grams per litre makes no allowance for variations in urine flow rate. Formal correction can be made by collecting timed urine samples and expressing results in protein mass per unit time. However, this is both inconvenient for the patient and prone to urine collection errors because of misunderstanding of when the collection should start and finish. Numerous studies have demonstrated that expression of urine total protein or urine albumin as a ratio to creatinine concentration in an early morning sample is at least as reproducible as a timed urine, even if collected under ideal conditions, and is more convenient for the patient. There are now few indications for collecting timed urine specimens.

Urine protein measurement

Throughout this chapter, it has been emphasized that pathologically significant proteinuria extends from the upper reference limit for urine protein, which is in milligrams per litre, to tens of grams per litre, in nephrotic syndrome. Chemical methods used for measuring urine total protein, which are both accurate and precise, are generally adequate for monitoring patients with proteinuria >1 g/L. However, there is a clear problem with using such techniques for quantitation of proteinuria in the mg/L range. First, external quality assessment programmes have highlighted the poor performance of urine total protein methods, particularly at low concentrations. This is partly related to the variable chromogenicity for the differing urine proteins of the different dyes used in total protein methods, resulting in different total urine protein reference ranges. Second, at low total protein concentrations, the relative contribution of proteins secreted into the urine by the kidneys, such as uromodulin, becomes more significant, thus masking within the total protein estimation small, but pathologically significant increases in urine albumin and other diagnostically important proteins (Fig. 8.2).

In the US National Kidney Foundation Guidelines, urine albumin is recommended as a sensitive marker for chronic kidney disease secondary to diabetes, glomerular disease and hypertension, and α_1 and β_2 -microglobulins as sensitive markers of tubulointerstitial disease. For this reason, many laboratories now provide urine albumin as their first line test for assessing 'proteinuria', covering the range from normal values, through microalbuminuria

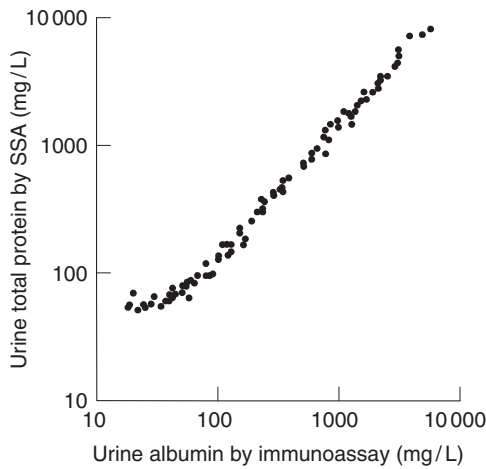


FIGURE 8.2 ■ Comparison of urine total protein measured by sulphosalicylic acid precipitation (SSA) with urine albumin measured by specific immunoassay. When proteinuria is in the g/L range, there is reasonably good agreement between methods, but at total protein concentrations within the reference range (<150 mg/L), pathological albuminuria (i.e. >20 mg/L) can be found. (Courtesy of Dr Robert Beetham).

to the albuminuric nephrotic range. However, at higher concentrations, the requirement for repeated dilutions of the sample for accurate measurement of albumin adds considerably to the cost. Above 0.5–1 g/L, some guidelines suggest that measurement of either urine albumin or total protein is acceptable for monitoring. In more

specialized circumstances, such as screening for urinary light chains or monitoring of tubular dysfunction, specific protein immunoassays are the most reliable.

Stepwise investigation of proteinuria

Where proteinuria is confirmed, a stepwise approach to establishing the cause of proteinuria can be applied, as summarized in Figure 8.3. Particular regard should be given to evidence of oedema, diabetes mellitus, polycystic disease, hypertension, heart failure and other cardiovascular risk factors. A full clinical and laboratory assessment at this stage will identify the cause of proteinuria in the majority of patients. For example, proteinuria may be caused by protein leakage from inflamed epithelium during urinary tract infections. In other patients, proteinuria, especially microalbuminuria, may be associated with non-renal conditions, some of which are transient and disappear with the resolution of the associated condition (Box 8.4).

A third group of patients will demonstrate persistent asymptomatic proteinuria either in isolation or accompanied by other urinary abnormalities. Red cells can enter urine or filtrate at any point along the urinary tract, leading to haematuria. Proteinuria with haematuria or tubular cell, granular or leukocyte casts is likely to originate from the nephron. When red cell casts are also found, the nephron can be implicated with more confidence and, in general, this indicates more severe renal disease.

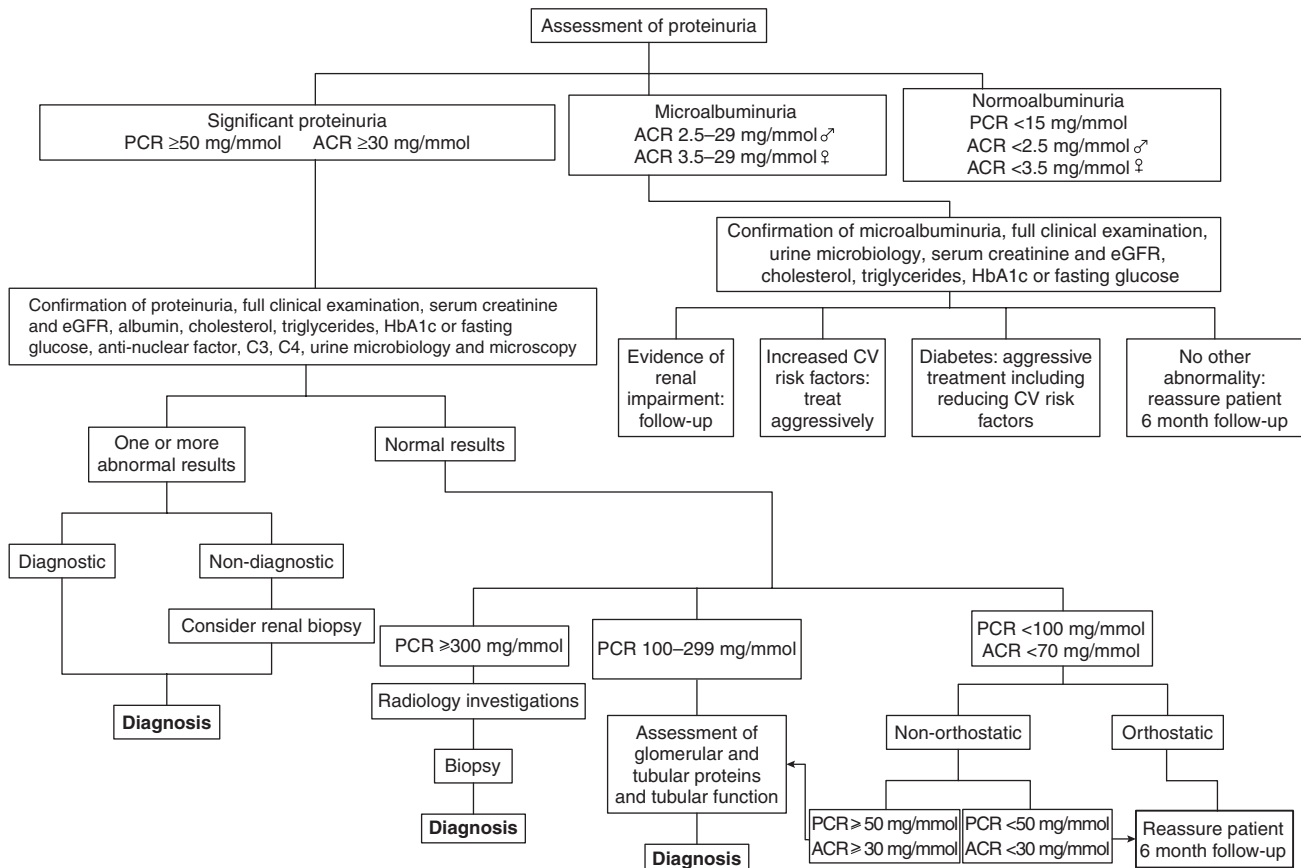


FIGURE 8.3 ■ Stepwise investigation of patients with proteinuria. Reference ranges and clinical decision points reflect current guidelines. PCR, urine protein creatinine ratio; ACR, urine albumin ratio; CV, cardiovascular.

BOX 8.4 Non-renal inflammatory conditions associated with microalbuminuria

- Trauma
- Burn injury
- Surgery
- Inflammatory conditions
 - Acute pancreatitis
 - Inflammatory bowel disease
 - Septicaemia
 - Systemic inflammatory response syndrome
- Cardiovascular disease (see text for details)
 - Myocardial infarction
 - Congestive cardiac failure
 - Intermittent claudication (especially post-exercise)
- Poisoning
 - Paracetamol (acetaminophen) overdosage

Studies have shown that ~70% of patients with asymptomatic non-orthostatic proteinuria have abnormal renal biopsies, while of those with orthostatic asymptomatic proteinuria, a much smaller proportion have abnormal biopsies. As a result, biopsy is now rarely performed in this latter group. The significance of the histological abnormalities is uncertain and the prognosis for both groups is excellent.

ACKNOWLEDGEMENT

I would like to thank Peter Gosling, who wrote this chapter for previous editions of the book.

Further reading

Although guidelines are available in print form, the latest versions incorporating updates and corrections are usually available only electronically (web addresses Accessed October 2013). Their titles are self-explanatory.

Chronic Kidney Disease Prognosis Consortium. Associated estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. *Lancet* 2010;375:2073–81.

A meta-analysis of the current evidence linking albuminuria and dipstick proteinuria with mortality and CKD.

Gerstein HC, Mann JF, Yi Q et al. Albuminuria and risk of cardiovascular events, death, and heart failure in diabetic and nondiabetic individuals. *J Am Med Assoc* 2001;286:421–6.

This was one of the largest prospective studies of microalbuminuria against outcome for a median of 4.5 years in 9000 diabetic and non-diabetic subjects. The results showed that any degree of albuminuria is a risk factor for cardiovascular events in diabetic and non-diabetic subjects, with the risk starting to increase well below the microalbuminuria cut-off. Screening for albuminuria identifies people at high risk for cardiovascular events. Numerous subsequent publications have served to confirm these findings.

James MT, Hemmelgarn BR, Wiebe N et al. for the Alberta Kidney Disease Network. Glomerular filtration rate, proteinuria and the incidence and consequences of acute kidney injury: a cohort study. *Lancet* 2010;376:2096–103.

The power of databases – A study of over 900 000 non-hospitalized patients that demonstrates the association of proteinuria as well as CKD with the risk of acute kidney injury.

Kidney Disease Improving Global Outcomes. <http://www.kdigo.org/home/guidelines>.

National Collaborating Centre for Chronic Conditions. Chronic kidney disease: national clinical guideline for early identification and management in adults in primary and secondary care. London: Royal College of Physicians; September 2008.

National Collaborating Centre for Chronic Conditions. Type 2 diabetes: national clinical guideline for management in primary and secondary care (update). London: Royal College of Physicians; 2008.

Newman DJ, Mattock MB, Dawney ABS et al. Systematic review on urine albumin testing for early detection of diabetic complications. *Health Technol Assess* 2005;9:30.

A useful review stating that regression of microalbuminuria to normoalbuminuria in adults is just as common as progression of microalbuminuria to clinical proteinuria; in children, regression is three-fold more frequent than progression.

NICE. <http://www.nice.org.uk/nicemedia/live/12069/42116/42116.pdf>.

NICE. <http://www.nice.org.uk/nicemedia/live/11983/40803/40803.pdf>.

NICE clinical guideline CG62: Antenatal care: routine care of the healthy pregnant woman (reviewed 2011), <http://guidance.nice.org.uk/CG62>.

NICE clinical guideline CG107: the management of hypertensive disorders during pregnancy; 2010. <http://guidance.nice.org.uk/CG107>.

NICE clinical guideline CG73: chronic kidney disease; 2008. <http://guidance.nice.org.uk/CG73>.

Rampoldi L, Scolari F, Amoroso A et al. The rediscovery of uromodulin (Tamm–Horsfall glycoprotein): from tubulointerstitial nephropathy to chronic kidney disease. *Kidney Int* 2011;80:338–47.

A useful recent review of current knowledge about uromodulin.

Renal Association. Clinical practice guideline on detection, monitoring and care of people with CKD. 5th ed; 2011. <http://www.renal.org/Clinical/GuidelinesSection/Detection-Monitoring-and-Care-of-Patients-with-CKD.aspx>.

Tojo A, Kinugasa S. Mechanisms of glomerular albumin filtration and tubular reabsorption. (Open access article) *International J Nephrol* 2012;2012:481–520.

A good overview and freely available.

Renal tubular disorders and renal stone disease

David Makanjuola • Marta Lapsley

CHAPTER OUTLINE

INTRODUCTION 168

RENAL TUBULAR DISORDERS 168

Introduction 168

Physiology 168

Isolated abnormalities of tubular function 169

Generalized tubular defects (Fanconi syndrome) 174

RENAL CALCULI 174

Introduction 174

Pathogenesis of renal stones 175

Investigation of stone formers 178

Treatment 178

CONCLUSION 179

APPENDIX 179

INTRODUCTION

Most patients with renal disease have some element of renal tubular involvement, but the other manifestations of the disease tend to be clinically more obvious and important. However, in a small number of patients, the clinical picture results primarily from a disorder of renal tubular function. These disorders can be inherited or acquired, and can affect tubular handling of a limited number of specific substances or encompass more generalized defects.

Renal tubular defects are conveniently considered with renal stone formation, since calculi sometimes form as a result of one of these conditions.

RENAL TUBULAR DISORDERS

Introduction

Hereditary renal tubular disease includes certain developmental disorders of the tubules, for example polycystic renal disease and medullary cystic disease. While these can result in disorders of renal function, including renal tubular function, they will not be considered in detail here. Renal tubular physiology will be discussed briefly, followed by a discussion of some well-recognized functional disorders of the renal tubules.

Physiology

Renal function is discussed in detail in Chapter 7, but in essence the process involves filtration at the glomeruli, followed by modification of this glomerular filtrate by

both tubular reabsorption and tubular secretion. Since 170L of filtrate are formed each 24h, but only about one-hundredth this amount of urine is produced, reabsorption is quantitatively the more significant (Table 9.1). This is largely an active, energy-requiring process and explains why the kidneys account for some 6–8% of the resting oxygen consumption of the body, while representing <1% of body mass.

Some of the mechanisms by which active transport in the renal tubules occurs are shown in Figure 9.1. The control of renal tubular handling of certain substances is covered in detail in other chapters, for example sodium and water in Chapter 4. Only the renal tubular handling of substances that are important in disorders of renal tubular function will be considered further here.

Glucose is absorbed with sodium ions in the early part of the proximal tubules, in a secondary active transport process. Glucose and sodium bind to a common carrier protein (SGLT 2, see later) in the luminal membrane and sodium moves down its electrochemical gradient, carrying glucose into the cell. Na^+, K^+ -ATPase in the non-luminal (basolateral) membrane of the tubular cells pumps the sodium ions out into the interstitial fluid, while glucose is transported in the same direction by the glucose transporter GLUT 2.

Amino acids are also reabsorbed in the early part of the proximal renal tubules, again by a secondary active transport system linked to sodium reabsorption. There appear to be separate cotransporter proteins for certain groups of amino acids, although some of these probably have overlapping specificities. The process is driven by the Na^+, K^+ -ATPase in the basolateral membrane pumping sodium out of the cell, with amino acids leaving by passive or facilitated diffusion.

TABLE 9.1 Renal tubular handling of various plasma constituents

Plasma constituent	Plasma conc. (mmol/L)	Filtered	Reabsorbed/24 h (mmol)	Secreted	Excreted in the urine
Sodium	140	23800	23700	–	100
Potassium	4.0	680	650	30	60
Bicarbonate	24	4080	4080	–	0
Urea	4.0	680	320	–	360
Creatinine	0.09	15.3	–	1.0	16.3
Urate	0.3	51	50	4	5
Glucose	4.5	765	765	–	0

Figures are illustrative only, assuming 170L filtrate is formed and reduced to 1.7L of urine in an adult on a normal diet. Note, while reabsorption is quantitatively more significant, some substances are secreted into the tubules.

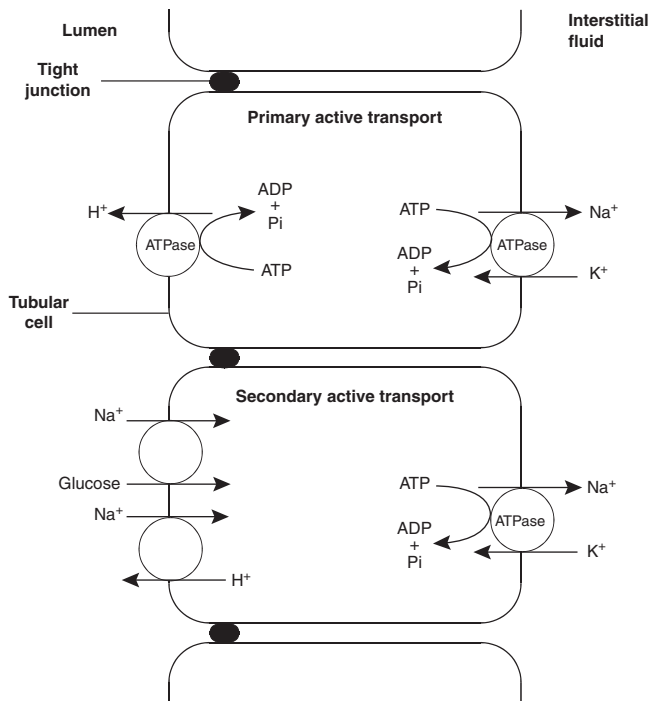


FIGURE 9.1 ■ Active transport mechanisms in the renal tubule.

Phosphate reabsorption in the renal tubules is influenced by the dietary intake of phosphate, certain hormones and a variety of other factors, and these are described in Chapter 6. In summary, about 90% of the inorganic phosphate in the plasma is freely filtered at the glomeruli, and then about 75% is reabsorbed in the proximal tubules. A small, variable amount is also absorbed in the distal tubules, but overall reabsorption is incomplete and up to 40 mmol/24h appears in normal adult urine. The rate-limiting step in reabsorption appears to be a secondary active transport system linked to sodium reabsorption, with a phosphate/sodium cotransporter located in the luminal membrane of the tubular cells.

As described in Chapter 5, the renal tubular secretion of hydrogen ions is linked to the net effective reabsorption of bicarbonate (see Figs 5.2 and 5.3). Around 4000 mmol of bicarbonate is filtered every 24h, but normal urine contains virtually no bicarbonate so the tubules must secrete 4000 mmol of hydrogen ions to achieve this. Although this mechanism prevents the loss of alkali into

the urine, it does not result in the net excretion of acid. The tubules must also secrete the hydrogen ions produced each day by normal metabolism (see Chapter 5), a further 40–80 mmol/24h.

There are two distinct mechanisms by which hydrogen ions are secreted into the tubular lumen (Fig. 9.1). A secondary active transport system linked to sodium operates in the epithelial cells of the early tubular segments, so that the Na^+, K^+ -ATPase on the basolateral membrane produces an electrochemical gradient for sodium to enter the cell from the luminal surface, but in contrast to glucose and amino acids, a hydrogen ion is simultaneously secreted into the lumen. Although a very high hydrogen ion gradient cannot be achieved, this mechanism is responsible for the bulk of hydrogen ion secretion, so that most bicarbonate reabsorption occurs in the proximal tubules. The sodium–bicarbonate cotransporter, located on the basolateral membrane of these tubular cells, mediates the transport of the generated bicarbonate into the systemic circulation. There may be other hydrogen ion secretory mechanisms in the proximal tubules, but they do not appear to be quantitatively important.

In the late tubular segments (the late distal tubules and the collecting ducts), a completely different mechanism for hydrogen ion secretion exists. This is relatively independent of tubular sodium content and occurs through primary active transport. The intercalated cells (I cells) in this part of the nephron have a hydrogen ion-transporting ATPase on their luminal surfaces and, although this accounts for <5% of the total hydrogen ions secreted, it is important because it can generate a hydrogen ion gradient of almost 1000 to 1. It is this that is responsible for the final acidification of urine and dictates the minimum achievable urinary pH of about 4.5 (or maximum hydrogen ion concentration of $\sim 32 \mu\text{mol/L}$). A chloride–bicarbonate exchanger at the basolateral membrane of the intercalated tubular cells is responsible for transporting the bicarbonate generated during this process to the systemic circulation.

Isolated abnormalities of tubular function

Glycosuria

Glucose is freely filtered at the glomeruli, but is normally then reabsorbed in the proximal tubules so that it is undetectable in urine. If plasma glucose concentrations rise to greater than $\sim 10 \text{ mmol/L}$ or if the glomerular filtration

rate increases (as in pregnancy), then the capacity of the proximal tubules to reabsorb filtered glucose is exceeded and glycosuria occurs.

Generalized defects in renal tubular function may also result in glycosuria (see later), but a small group of people appear to have an isolated defect of tubular glucose reabsorption.

Hereditary renal glycosuria. Patients with this condition excrete a variable amount of glucose in their urine at normal plasma glucose concentrations. Other aspects of carbohydrate metabolism are not affected: glucose tolerance and plasma insulin concentrations are normal. The condition is inherited in an autosomal recessive manner, and two major phenotypes have been identified (types A and B), based on the exact changes in the kinetics of glucose reabsorption. Hereditary renal glycosuria is rare, and is generally recognized to be a benign condition with no clinical sequelae.

The reabsorption of glucose in the proximal renal tubules is similar to the absorption of glucose in the intestine. There is a Na^+ ,D-glucose cotransporter on the luminal cell wall to transport glucose into the cell, with a facilitated glucose transporter (GLUT 2) on the basolateral membrane to enable the glucose to exit. The intestinal form of the Na^+ ,D-glucose cotransporter (SGLT 1) and its gene have been well characterized, but, while the corresponding renal cotransporter (SGLT 2) is known to differ from SGLT 1, its identity in humans is less well established. Nevertheless, it is assumed that mutations in the gene for SGLT 2 cause hereditary renal glycosuria by interfering with the uptake of glucose in the proximal tubules.

There is a corresponding condition affecting SGLT 1 in the gut. This cotransporter is involved in both glucose and galactose absorption, and the classic presentation is with life-threatening diarrhoea in early infancy due to glucose and galactose malabsorption (familial glucose-galactose malabsorption). There is often an associated mild renal glycosuria, although in hereditary renal glycosuria there is no corresponding effect on the gut.

Mutations affecting GLUT 2, which facilitates transport of glucose, galactose and fructose across the basolateral membrane, are another rare cause of renal glycosuria (Fanconi-Bickel syndrome).

Amino acidurias

Amino acids are normally freely filtered at the glomeruli and then almost entirely reabsorbed in the proximal convoluted tubules. There is a maximal capacity to each reabsorptive mechanism and, in most patients with amino aciduria, some extrarenal disorder leads to accumulation of amino acid(s) in the plasma, that exceeds the reabsorptive capacity of the tubules, with consequent 'overflow' amino aciduria. However, as with glycosuria, generalized defects in renal tubular function can result in amino aciduria, and there are also some isolated defects in the reabsorption of particular groups of amino acids.

Cystinuria. Cystinuria is the classic example of an amino aciduria due to a defect in renal tubular function, in that the amino aciduria occurs at normal or even low plasma concentrations of the amino acids involved.

In most patients with cystinuria, there is renal loss not only of cystine, but also of the dibasic amino acids ornithine, arginine and lysine. There is also an associated failure of intestinal absorption of the same amino acids. Inspection of their molecular structures (Fig. 9.2) shows that each has two amino groups separated by 5–7 bonds, which suggests that malfunction of a single membrane carrier protein might explain the disorder. However, the true explanation is not this simple, since the clearance of cystine may exceed the creatinine clearance, suggesting secretion of cystine into the tubules and, furthermore, since dibasic amino aciduria (e.g. lysinuric protein intolerance) or cystinuria can each occasionally occur alone.

The only known clinical manifestation of cystinuria is recurrent urinary tract stone formation, the name 'cystine' coming from the original (erroneous) assumption that the source of these stones was the bladder. Cystine stones form readily in acidic urine. They are yellow-brown in colour and are radio-opaque because of their sulphur content, although they are less radio-dense than calcium-containing stones. There may also be some calcium deposition if there is infection secondary to the calculi. They tend to occur as staghorn or multiple recurrent stones and often require some form of surgical intervention or lithotripsy (fragmentation of stones by external shock waves). Patients with cystinuria also have a higher incidence of calcium oxalate stones than the general population. All stone formers should,

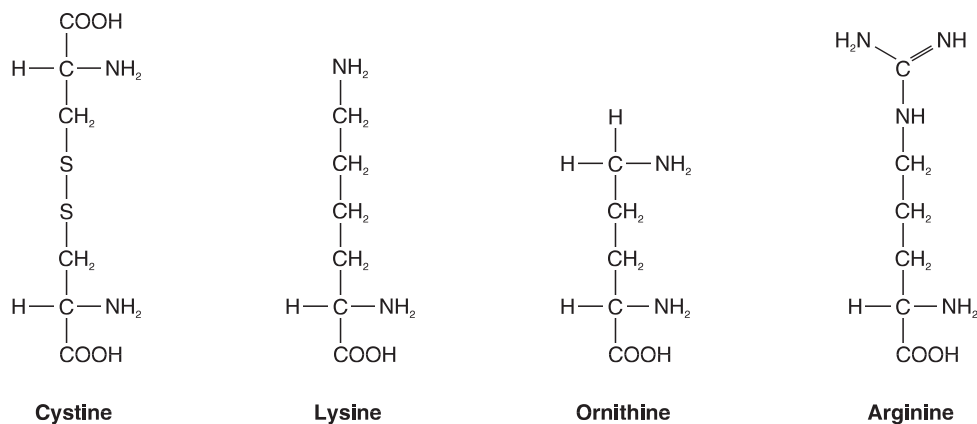


FIGURE 9.2 ■ Chemical structures of the dibasic amino acids involved in cystinuria.

therefore, be screened for cystinuria, preferably by formal amino acid measurement in the urine, as qualitative screening tests are not sufficiently sensitive. The intestinal defect appears to cause no clinical problems.

Cystinuria occurs with equal frequency in both sexes, although males tend to be more severely affected. It may present at any time from the first year of life up to the ninth decade, with a peak incidence in the second and third decades. The prevalence of cystinuria varies between racial groups and according to whether the figures are taken from neonatal amino acid screening programmes or from known cystinuric stone formers (immaturity of the renal tubules in the first few months of life may lead to some heterozygous infants having urinary cystine outputs in the homozygous cystinuria range, leading to misclassification). However, the worldwide prevalence is estimated to be 1 in 7000, making it a relatively common inherited metabolic disorder.

The mode of inheritance is autosomal recessive, although in some families it appears to be incompletely recessive, with heterozygotes excreting more urinary cystine, ornithine, arginine and lysine than normal, although less than in the homozygous state. In the past, this has led to attempts to classify cystinuria into type I (homozygous for the fully recessive form) and types II and III (variants of the incompletely recessive form), based on combinations of the relative amounts of the amino acids excreted in the urine and/or the intestinal uptake of cystine and dibasic amino acids. The discovery of the genes involved in the recessive and the incompletely recessive forms of cystinuria has made this classification clearer, although the situation is not yet completely resolved.

The gene involved in type I cystinuria codes for a protein known as rBAT (a loose acronym for 'related to the b⁰⁺ amino acid transporter', with the term b⁰⁺ signifying broad specificity for neutral (0) and dibasic (+) amino acids). rBAT is a membrane glycoprotein that is one of the activating components of a heteromultimeric transporter for cystine and dibasic amino acids. The gene for rBAT (*SLC3A1*) is located at 2p16.3-p21, and, to date, more than 120 different mutations have been identified in it in patients with type I cystinuria.

A second cystinuria locus, accounting for non-type I forms of cystinuria, has been identified by linkage analysis at chromosome 19q13.1–13.2. This gene (*SLC7A9*) encodes a protein that acts as the amino acid transporting subunit of the membrane complex. More than 95 gene mutations have been identified. Most, but not all cases of cystinuria can be explained by mutations in *SLC3A1* and *SLC7A9*.

Healthy individuals excrete <10 μmol cystine/mmol of creatinine (<130 μmol/24h). Patients with homozygous cystinuria usually excrete >1700 μmol/24h and may

excrete up to 5000 μmol/24h. Heterozygotes may have an entirely normal amino acid excretion pattern in the urine or may excrete up to 1700 μmol cystine/24h, in which case they may produce stones.

Medical treatment of cystinuria begins with maintenance of a high fluid intake throughout both day and night and alkalization of the urine, both being aimed at decreasing the chance of precipitation of cystine in the renal tract. The solubility of cystine is pH dependent with a solubility limit of ~700 μmol/L at pH7, rising to ~1500 μmol/L at pH7.5. Regular monitoring is required to ensure that cystine remains well below its solubility limit throughout the day and, critically, the night. If these measures fail, then it may be possible to convert the cystine to a more soluble compound, most commonly by the use of D-penicillamine. This can form the mixed disulphide cysteine-penicillamine (Fig. 9.3), which is significantly more soluble than cystine. The aim is to keep the daily excretion of free cystine below 2000 μmol. Unfortunately, D-penicillamine frequently causes an allergic reaction and can also cause nephrotic syndrome and pancytopenia, so careful monitoring is essential. Second generation chelating agents include tiopronin (α-mercaptopropionylglycine) and captopril, although studies on the efficacy of the latter have been inconclusive. Small stones may dissolve with careful medical management, although frequent surgical intervention is required for some patients.

Occasionally, in spite of both medical and surgical treatment for stone formation, cystinuria causes sufficient renal damage to result in established renal failure. In this case, renal transplantation may be effective, since the donor kidney should not be affected by the amino acid transport defect and the patient should, therefore, remain disease free.

Hartnup disorder. A second example of an amino aciduria due to a true defect in renal tubular function, rather than an 'overflow' effect, is found in Hartnup disorder. This is named after the family in which it was first described and is again a defect of both renal and intestinal amino acid transport. The constant feature is a failure to reabsorb the neutral amino acids (Box 9.1) in the renal tubules, with their consequent appearance in the urine. The failure of reabsorption is not absolute, as renal clearances of the affected amino acids are generally lower than the creatinine clearance. Most affected individuals also have increased amounts of indoles (e.g. indican) in the urine, which originate from the bacterial breakdown of unabsorbed tryptophan in the gut. Rarely, the renal or intestinal lesions may occur alone.

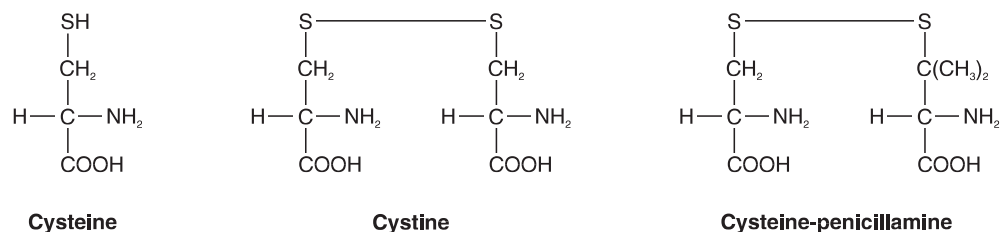


FIGURE 9.3 ■ Chemical structures of cysteine, cystine and cysteine-penicillamine.

BOX 9.1 The neutral amino acids that appear in the urine in excess in Hartnup disorder

Excretion of the neutral sulphur containing amino acid methionine may also be increased.

Monoamino-monocarboxylic amino acids

- Alanine
- Serine
- Threonine
- Valine
- Leucine
- Isoleucine
- Phenylalanine
- Tyrosine
- Tryptophan
- Histidine

Monoamino-dicarboxylic amides

- Asparagine
- Glutamine

The original description of Hartnup disease included a pellagra-like skin rash, transient cerebellar ataxia and constant renal amino aciduria. Some affected individuals have also shown psychotic behaviour, while others have learning disabilities. The pellagra-like rash and its response to nicotinamide suggest that the clinical features of the disease may be due to failure to absorb tryptophan in the intestine and reabsorb it in the renal tubules, leading to a deficiency of nicotinamide. However, investigation of siblings of individuals with Hartnup disease and the results of neonatal urine amino acid screening programmes both suggest that the typical amino aciduria (Hartnup disorder) may exist without the features of Hartnup disease. It would appear that Hartnup disorder is an autosomal recessive inherited condition, but that this is not expressed clinically without the presence of certain other genetic or environmental influences, such as poor nutrition. The causative gene, *SLC6A19*, is located on chromosome 5p15.33 and codes for a sodium-dependent neutral amino acid transporter.

Familial renal iminoglycinuria. A further example of a specific abnormality in renal tubular reabsorption causing a distinct pattern of amino/imino aciduria occurs in familial renal iminoglycinuria. The primary genetic defect is a mutation in either the *SLC36A2* gene, which codes for a high affinity glycine, proline and hydroxyproline transporter, or the *SLC36A1* gene, which codes for a low affinity transporter. However, the phenotype is variable as it may be modified by mutations in at least three other genes. Some reabsorption of these amino acids can still occur via the transport system that is not affected in any individual patient.

The condition, which is benign, is autosomal recessive, although some heterozygotes are 'incomplete' and have hyperglycinuria. In some but not all homozygotes, impaired intestinal transport of proline can be demonstrated.

Neonatal screening for amino acidurias suggests that familial renal iminoglycinuria occurs in 1 in 15000 live

births in a Caucasian population. However, it should be appreciated that in healthy neonates, the renal tubular reabsorption of proline, hydroxyproline and glycine is less efficient than in adults; imino aciduria normally disappears by three months and hyperglycinuria by six months of age.

Dent disease

Also known as X-linked nephrolithiasis, Dent disease is most often caused by a mutation in the gene coding for a voltage gated chloride channel (CLC-5) expressed predominantly in the proximal renal tubules. However, a significant number of patients have a mutation in the gene usually associated with the oculocerebrorenal syndrome of Lowe, *OCRL*. Both defects are thought to cause failure of the endosomal acidification that is required to allow recycling of the proximal tubular membrane receptor, megalin, which mediates protein reabsorption. The syndrome is characterized by hypercalciuria, nephrocalcinosis, recurrent calcium-containing renal stones, low molecular weight proteinuria and chronic kidney disease eventually leading to established renal failure. Glycosuria, phosphaturia, amino aciduria and rickets may also occur in affected males. Carrier females rarely present clinically, but often have increased urinary excretion of low molecular weight proteins. Mutations of the gene encoding another voltage gated chloride channel (CLC-Kb) are responsible for some cases of Bartter syndrome (see p. 57).

Phosphate transport defects

Specific disorders of renal tubular phosphate handling can be inherited or acquired. There is reduced proximal tubular reabsorption of phosphate (in particular, the TmP/GFR is reduced, as explained in Chapter 6) and hypophosphataemia. The normal response to this would be an increase in the 1 α -hydroxylation of vitamin D, and if this is adequate, the end result would be hypercalciuria (with an increased risk of renal stones), but no bone disease. If the vitamin D response is inadequate, then the result may be bone disease with hypercalciuria (e.g. as in hereditary hypophosphataemic rickets with hypercalciuria) or bone disease alone (e.g. X-linked hypophosphataemic rickets and oncogenic hypophosphataemic osteomalacia). Inherited and acquired forms of hypophosphataemic bone disease are discussed further in Chapter 31.

The phosphate loss, and other changes in more generalized forms of renal tubular disorder, may also be severe enough to cause bone disease (see p. 174).

Renal tubular acidosis

The renal tubular acidoses (RTAs) are a group of hyperchloraemic metabolic acidoses that occur secondarily to an abnormality in urinary acidification. The plasma bicarbonate concentration is low, the plasma chloride raised and the anion gap normal, in contrast to the acidosis seen with reduction in the glomerular filtration rate in which the plasma bicarbonate is low, chloride normal and anion gap increased.

Renal tubular acidoses are classified according to the presumed site of the failure of hydrogen ion secretion. In proximal (type 2) RTA, there is a failure in bicarbonate reabsorption, while in distal RTA, there is a failure in net acid excretion. If failure of hydrogen ion secretion is the sole defect in the distal nephron, it is known as classic distal (type 1) RTA, whereas if the defect is more generalized, causing hyperkalaemia as well as acidosis, it is known as generalized distal (type 4) RTA. These conditions are described in Chapter 5, but the underlying mechanisms are briefly summarized again here, and details of diagnostic tests are given in [Appendix 9.1](#). Type 3 RTA is a mixture of types 1 and 2 and is very rare. The term was originally used to describe a transient condition occurring in infants with immature tubules. However, it is now usually applied to a rare inherited deficiency of carbonic dehydratase type II, in which infants also have cerebral calcification, mental retardation and osteopetrosis (see Chapter 31).

Proximal (type 2) renal tubular acidosis. Isolated proximal RTA may be autosomal recessive or dominant. A mutation in the sodium–bicarbonate cotransporter gene (*SLC4A4*) is responsible for the recessive form; the molecular basis for the dominant form is yet to be elucidated. The net effect is that bicarbonate reabsorption is impaired, leading to metabolic acidosis and bicarbonaturia, with a decrease in the plasma bicarbonate concentration until the filtered load of bicarbonate falls to a point at which hydrogen ion secretion is sufficient to allow reabsorption of all filtered bicarbonate. Thus, a new steady state is established, in which there is a metabolic acidosis with a low plasma bicarbonate concentration (15–20 mmol/L) but no bicarbonaturia.

Administration of bicarbonate to such patients will establish a steady state closer to normal, but quite large doses of bicarbonate are required (e.g. 5–15 mmol/kg body weight/24h), since as the plasma bicarbonate rises, so does the magnitude of renal bicarbonate loss. Hydrogen ion secretion in the distal tubules is normal and the urine can be acidified in response to an acid load.

Proximal RTA may occur as an isolated defect, but is more usually part of a generalized defect of proximal tubular function, the Fanconi syndrome (see [p. 174](#)).

Distal (type 1) renal tubular acidosis. Decreased activity of the apical H^+ -ATPase (proton pump) is probably the commonest defect in distal RTA and may be the end result of a number of different pathological processes. The most common of these is Sjögren syndrome, which should be excluded in any patient presenting in adulthood. Other dysproteinaemic and autoimmune conditions may also cause acquired distal RTA, as may chronic hypercalciuria with nephrocalcinosis or treatment with drugs such as amphotericin, lithium and analgesics. Inherited forms (which may be autosomal dominant or recessive) include mutations in the genes encoding the H^+ -ATPase pump and the chloride–bicarbonate exchanger. There is a complex relationship between distal RTA, nephrocalcinosis and kidney stone formation. In some patients, hypercalciuria appears to be the primary abnormality, causing interstitial and tubular damage with calcium deposition and secondary

RTA. In others, RTA is the primary condition causing increased calcium phosphate resorption from bone in an attempt to buffer the retained acid; the subsequent hypercalciuria is proportional to the severity of the acidosis.

While other distal tubular functions remain unimpaired, hydrogen ion secretion is reduced so that urinary pH cannot be lowered below 5.5. A resultant increase in the exchange of potassium for sodium ions in the distal tubules causes significant renal potassium loss and the condition is usually accompanied by hypokalaemia.

Diagnosis depends on the demonstration of an inappropriately high urinary pH (>5.5) during either spontaneous acidosis or an acid loading test (see [Appendix 9.1](#)). Treatment involves potassium supplementation together with bicarbonate. Unlike proximal RTA, the dose of bicarbonate required is small (e.g. a little over 1 mmol/kg body weight/24h in adults), since there is only a minor degree of bicarbonaturia and this does not increase as the plasma bicarbonate concentration rises.

Patients with incomplete distal RTA have low urinary citrate excretion and urine pH persistently >5.5 but are able to maintain their plasma bicarbonate concentrations within the reference range, possibly by increasing proximal tubular ammonium production. They may also have hypercalciuria and are at risk of developing osteoporosis, nephrocalcinosis and calcium phosphate kidney stones. Hypercalciuria may be reversed and the risk of developing osteoporosis and kidney stones reduced by alkali therapy, which decreases calcium phosphate resorption from bone.

Distal renal tubular acidosis with hyperkalaemia (type 4). In type 4 RTA, there is a generalized impairment of distal tubular function associated with aldosterone deficiency or resistance, with decreased tubular secretion of both hydrogen and potassium ions. A relatively severe hyperkalaemia is usually the primary abnormality and the accompanying hyperchloraemic metabolic acidosis is often mild with a plasma bicarbonate concentration >16 mmol/L.

The condition may arise because of a lack of mineralocorticoid activity (e.g. in adrenal failure), disease of the kidney resulting in impaired production of renin (hyporeninaemic hypoaldosteronism, e.g. in diabetic nephropathy) or resistance to the action of mineralocorticoids on the distal tubules (e.g. in patients treated with spironolactone). Treatment of the condition depends on the underlying cause.

Hereditary renal hypouricaemia

This is a rare disorder of the renal tubular handling of urate, in which there is a net decrease in reabsorption of urate, resulting in hypouricaemia and an increased renal urate clearance. Plasma urate concentrations are <150 $\mu\text{mol/L}$ in men and <126 $\mu\text{mol/L}$ in women. The condition is generally harmless, although some patients also have hypercalciuria, and about 25% have a tendency to stone formation. It is transmitted in an autosomal recessive manner and heterozygotes tend to have intermediate plasma urate concentrations. Mutations have been demonstrated in the gene for a urate–anion exchanger, which is found on the luminal membrane of proximal

tubular cells and which accounts for the majority of urate reabsorption in normal kidneys.

Generalized tubular defects (Fanconi syndrome)

Generalized renal tubular defects tend to occur together in a distinct syndrome known as the renal Fanconi syndrome (which should not be confused with Fanconi anaemia, a form of congenital aplastic anaemia or Fanconi–Bickel syndrome, a rare disorder of carbohydrate metabolism). There is a failure in net proximal tubular reabsorption of glucose, amino acids, phosphate and bicarbonate, with consequent glycosuria, amino aciduria, phosphaturia and acidosis, together with the development of vitamin D-resistant metabolic bone disease. There may also be increased urinary losses of water and other substances, for example sodium, potassium, calcium, magnesium, urate and low molecular weight proteins. The exact mechanism by which the Fanconi syndrome occurs is not clear and there may be distal, as well as proximal, tubular dysfunction.

Causes of the renal Fanconi syndrome may broadly be divided into inherited and acquired and are listed in Box 9.2. The most common cause of inherited Fanconi syndrome in children is cystinosis, an autosomal recessive disorder of cystine transport across lysosomal membranes. The gene responsible (*CTNS*) is located on chromosome 17p13 and encodes a lysosomal membrane protein, cystinosin. There is accumulation of cystine in the lysosomes of most tissues, including the kidneys, and in the infantile form of the disease, the Fanconi syndrome progresses to cause glomerular damage and renal failure during childhood. Cystinosis is usually treated with cysteamine, which is concentrated inside the lysosomes and reacts with cystine to form both cysteine and a cysteine–cysteamine complex. These are able to cross the lysosomal membrane via different transporters. When administered regularly, cysteamine decreases the cystine

load in lysosomes, conserves renal function and improves growth. The adequacy of treatment can be assessed by measuring the cystine content of white blood cells. If renal function becomes significantly impaired, transplantation is particularly useful, since the transplanted kidney does not have the genetic defect.

There is also an adult onset form of cystinosis in which there is no Fanconi syndrome or glomerular impairment, and an intermediate form that presents during adolescence and does eventually lead to renal failure.

Other inherited metabolic diseases that are associated with the Fanconi syndrome are listed in Box 9.2. There is also an idiopathic inherited form, but this is a diagnosis of exclusion. It is interesting that in inherited metabolic diseases for which there is a specific treatment (e.g. avoidance of lactose and galactose in galactosaemia), the Fanconi syndrome resolves on treatment, presumably because of the fall in the concentration of a toxic metabolite.

While the inherited diseases associated with the Fanconi syndrome tend to present in childhood, the acquired forms tend to occur in adults. These may also be reversible if exposure to the causative agent can be stopped.

The clinical features of the Fanconi syndrome tend to be rather non-specific and include polyuria, polydipsia, dehydration, hypokalaemia and acidosis, with impaired growth and rickets in children and osteomalacia in adults.

Treatment is primarily directed at the underlying cause, but inappropriate renal losses need to be replaced (e.g. with fluids, bicarbonate and potassium) and the bone disease treated with phosphate replacement and vitamin D.

RENAL CALCULI

Introduction

Stone formation in the urinary tract (urolithiasis) has been described since ancient times, but in the past, lower urinary tract stones, particularly those arising in the bladder, appear to have been more frequent, engendering a surgical enthusiasm for ‘cutting for stone’. In industrialized, relatively affluent populations, renal stone formation (nephrolithiasis) has increased in frequency, while bladder stone formation has almost disappeared as the prevalence of malnutrition and infection have decreased.

However, while the frequency of nephrolithiasis is acknowledged to be increasing in the developed world, precise incidence and prevalence figures have not been well established. Nevertheless, it would appear that as many as 5% of the population (possibly even more) will have a clinical stone event at some time during their lives. In previous decades, stone formation was more frequent in men than in women, but in recent years the incidence has become almost equal. The ten-year risk of further stone recurrence can be as high as 70% after a first episode, so investigation and treatment, which can reduce this figure to 25%, is worthwhile.

The clinical effects of stone formation tend to be similar whatever the type, but it is apparent that there are a number of distinct metabolic derangements that can each

BOX 9.2 Some causes of renal Fanconi syndrome

Inherited

- Idiopathic
- Oculocerebrorenal (Lowe) syndrome
- Associated with inherited metabolic disease:
 - Cystinosis
 - Fructose intolerance
 - Galactosaemia
 - Glycogen storage disease type I
 - Tyrosinaemia
 - Wilson disease

Acquired

- Exogenous toxins
 - Heavy metals: lead, cadmium, mercury, copper
 - Drugs (e.g. tetracycline that has been stored for too long, gentamicin, ifosfamide, tenofovir)
- Paraproteinaemia
- Amyloidosis

TABLE 9.2 The main types of renal stone, in approximate order of frequency of occurrence

Type of stone	Frequency (%)
Predominantly calcium oxalate	60
Uric acid	17
Magnesium ammonium phosphate and calcium phosphate (struvite)	12
Predominantly calcium phosphate	10
Cystine	1

give rise to stones of a characteristic composition. The main types of renal stones are listed in Table 9.2 in approximate order of frequency of occurrence, and aetiological factors in the formation of each are described in detail below.

The classic clinical manifestation of nephrolithiasis is renal or ureteric colic. The pain is characteristically severe, requiring opiates for analgesia, and is usually accompanied by haematuria. Some stones may be discovered incidentally when abdominal X-rays are taken for an unrelated reason, and passage of tiny stones may cause only minimal discomfort. In general, permanent intrinsic renal damage rarely occurs after an acute event unless there is superadded infection in an obstructed kidney. However, there is an increased incidence of chronic kidney disease in patients with a history of kidney stones and current UK guidelines recommend screening such patients annually with measurement of plasma creatinine and calculation of eGFR (see Chapter 7).

Pathogenesis of renal stones

The role of the kidneys in water conservation means that it is often necessary to excrete concentrated urine. Since some constituents of urine are relatively insoluble in water, to the extent that supersaturated solutions form, it is not surprising that these constituents sometimes crystallize out to form renal calculi.

Certain factors predispose to this process: chronic dehydration, as may occur in hot climates, is one factor. Increased urinary excretion of certain constituents (e.g. calcium, urate or oxalate) increases the likelihood of supersaturation occurring, and alteration of urinary pH may adversely affect the solubility of some solutes. For example, urate stones are more likely to form in acidic urine, whereas alkaline conditions, as may occur with a urinary tract infection, make calcium phosphate precipitation more likely. There is evidence that specific microorganisms may precipitate early crystallization. Anatomical abnormalities of the kidneys, such as medullary sponge kidney or pelvi-ureteric junction obstruction, increase the risk of stone formation, at least in part because of urinary stasis. There is a 2–3 fold risk of stone formation in people with a family history of urolithiasis. Patients who are overweight, especially if they have insulin resistance, have an increased risk of developing both calcium oxalate and uric acid stones, partly because of excretion of excess dietary constituents in the urine and partly because of the production of persistently acidic urine.

These factors are opposed by the presence in urine of inhibitors of crystallization (e.g. magnesium, pyrophosphate, citrate and certain glycoproteins), so that crystal formation in urine is usually slower than in simple salt solutions.

Each of the types of stone listed in Table 9.2 will now be considered in more detail.

Calcium stones

Calcium-containing stones form the majority of renal calculi and may consist of calcium oxalate, calcium phosphate or a mixture of the two. Hyperoxaluria, with or without hypercalciuria, is likely to be involved in the formation of pure calcium oxalate stones, whereas pure calcium phosphate stones suggest either hypercalciuria or persistently alkaline urine.

Hypercalciuria. Hypercalciuria is present in around 30% of patients with calcium-containing stones, although the exact prevalence depends on the population studied and the upper limit accepted for the reference range (of the order of 7.5 mmol/24 h for men and 6.25 mmol/24 h for women or 0.1 mmol/kg body weight/24 h).

A minority of hypercalciuric stone formers also have hypercalcaemia, and it is then important to establish a cause for this; it is most likely to be due to primary hyperparathyroidism. The treatment is that of the underlying cause.

Normocalcaemic hypercalciuria may be divided broadly into four groups: skeletal resorption, dietary, absorptive and renal leak hypercalciuria. Skeletal resorption and hence urinary excretion of calcium may be seen in patients following prolonged immobilization, Paget disease of bone or excessive glucocorticoid activity including chronic steroid therapy. Pure dietary hypercalciuria is rare, but high consumption of dairy products, combined calcium and vitamin D supplements or salt all tend to increase urinary calcium excretion.

Increased intestinal absorption of calcium accounts for around 50% of subjects with normocalcaemic hypercalciuria. The mechanism is not well understood, but there is some evidence for increased end-organ sensitivity to vitamin D, resulting in absorption of a greater fraction of dietary calcium than in normal subjects. Plasma parathyroid hormone (PTH) concentrations may be decreased.

Renal leak hypercalciuria may be seen in renal tubular acidosis, medullary sponge kidney, chronic loop diuretic therapy and proximal tubular syndromes such as Dent disease, but is more often idiopathic, accounting for 25% of hypercalciuric subjects. Some patients may have an unrecognized bone resorption defect; indeed, it is not always possible to determine whether the primary disorder is of renal or bone origin. Patients may be differentiated from those with absorptive hypercalciuria by a calcium load test: calcium excretion is high after loading in both groups but remains >0.23 mmol/mmol creatinine after fasting in patients with a renal leak of calcium (see Appendix 6.1). Plasma PTH is often mildly raised in response to the chronic calcium loss, although any increase is more commonly due to vitamin D deficiency. Treatment with thiazide-like diuretics reduces both

calcium excretion and plasma PTH concentrations, and reduces the risk of osteoporosis.

Treatment of patients with hypercalciuria and/or calcium containing stones includes increasing fluid intake to achieve daily urine volumes of 2–3 L, correcting dietary factors such as excessive salt, protein, sugar and calorie intake, and avoiding combined calcium and vitamin D supplements. A low sodium diet enhances both sodium and calcium reabsorption in the proximal tubules and so decreases urinary excretion (calcium is reabsorbed passively along the concentration gradient produced by active sodium and water reabsorption). Restricting animal protein reduces the dietary acid load and therefore the loss of calcium from bone. Restricting sucrose and fructose consumption appears to reduce urinary calcium excretion and may help with weight loss, which reduces the risk of stone recurrence. Formal calcium restriction is no longer recommended owing to the risk of osteoporosis as well as to a paradoxical rise in stone risk from the increased absorption of free oxalate in the intestine (calcium oxalate is poorly absorbed). Subjects should be advised to consume a ‘normal’ amount of calcium per day of around 1000 mg. However, calcium supplements, especially if taken between meals, should be avoided as they are associated with an increased risk of stone formation. Thiazide-like diuretics, such as chlortalidone or indapamide, specifically reduce renal calcium excretion and are particularly successful in patients with renal leak hypercalciuria. Those with absorptive hypercalciuria may become refractory to treatment and require an occasional ‘thiazide holiday’. Response to treatment should be monitored by measuring urinary calcium excretion. Biochemical side-effects, including hypokalaemia, hyperuricaemia, hypercalcaemia and hyperglycaemia, are relatively common and should be sought routinely. Potassium citrate supplements should be added, especially if the urinary citrate excretion is below the mid range and the urine pH is regularly <6.5, or if there is hypokalaemia. Excessive treatment with potassium citrate, resulting in the production of alkaline urine, increases the risk of calcium phosphate stone formation and should be avoided. Patients with distal renal tubular acidosis may show a sufficient fall in urinary calcium excretion in response to potassium citrate alone to avoid treatment with thiazide diuretics. This group includes patients with a partial form of the disorder in which plasma pH is normal under basal conditions, but the urine cannot be acidified when tested formally.

Hyperoxaluria. Now that oxalate can be measured reliably in the urine, it has become apparent that hyperoxaluria contributes as much, if not more, to calcium oxalate stone formation as hypercalciuria. A relatively small increase in urinary oxalate excretion above normal can have a marked effect on the risk of stone recurrence. In most instances, hyperoxaluria results from excessive dietary intake or increased intestinal absorption of oxalate; the two rare primary hyperoxalurias are described below.

The upper reference limit for urinary oxalate excretion is of the order of 400–500 $\mu\text{mol}/24\text{h}$. A dietary excess of oxalate-containing foods can increase excretion to

700 $\mu\text{mol}/24\text{h}$ or even greater, particularly if the dietary calcium content is low.

Increased intestinal absorption of oxalate may occur in short bowel syndrome and malabsorption, whatever the cause (enteric hyperoxaluria). Unabsorbed fatty acids in the gut lumen combine with calcium, reducing the available calcium for calcium oxalate formation. This leaves larger than usual amounts of oxalate in the free form, which can be easily absorbed. In addition, exposure of the colonic mucosa to substances such as fatty acids and bile salts, which have detergent properties, may increase its permeability to oxalate and further increase the amount absorbed.

Primary hyperoxaluria is a rare inherited metabolic disease that must be considered if nephrolithiasis occurs during childhood (although a few affected individuals do not present until adult life). There is increased synthesis of oxalate, with a consequent increase in its urinary excretion, together with that of some other organic acids. The pattern of accompanying organic aciduria has allowed classification into two types. In type 1 primary hyperoxaluria, excessive amounts of glyoxylic and glycolic acids are excreted, whereas in the rarer type 2, excretion of these acids is normal but that of L-glyceric acid is increased. The biochemical defect in type 1 is a deficiency of the peroxisomal enzyme alanine:glyoxylate aminotransferase. In type 2, the defect is in a cytosolic enzyme, D-glycerate dehydrogenase/glyoxylate reductase. The inheritance of both types is autosomal recessive. In type 1, the 24 h excretion of oxalate is of the order of 1.5–3 mmol; it may be lower in type 2, which may have a less severe clinical course. Besides renal stone formation, there is also a tubulointerstitial nephropathy that progresses to chronic kidney disease. Plasma oxalate concentrations begin to rise as renal clearance decreases, eventually leading to widespread tissue deposition of oxalate (e.g. in myocardium, retinae and synovial membranes). However, the condition appears to exhibit marked clinical and biochemical heterogeneity.

Other causes of hyperoxaluria include ingestion of ethylene glycol, which is rapidly converted to glycolate and can lead to calcium oxalate crystals forming in the renal tubules, excessive ingestion of vitamin C, which may lead to increased oxalate formation in some individuals, and chronic *Aspergillus* infection. Pyridoxine is a cofactor for the alanine:glyoxylate aminotransferase enzyme, and deficiency results in mild hyperoxaluria that responds to physiological replacement doses. A syndrome of ‘mild metabolic hyperoxaluria’ has also been described in adults that responds to pharmacological doses of pyridoxine.

Treatment of calcium oxalate stone formers includes increased fluid intake, dietary restriction of oxalate (Box 9.3) and oral potassium citrate if urine citrate excretion is below mid range or the pH is consistently <6.5. If there is no hypercalciuria, oral calcium supplements can be taken at mealtimes to decrease intestinal oxalate absorption. The starting dose is 500 mg equivalent of calcium per day, which may be increased to 10 g/day in patients with severe malabsorption and should be titrated against urine calcium and oxalate excretion. Pyridoxine reduces urinary oxalate excretion in some individuals: the starting dose should be 10 mg/day, increasing to a

BOX 9.3 Foods rich in available oxalate

- Tea (especially black)
- Ovaltine
- Drinks rich in vitamin C (converted to oxalate)
- Spinach
- Rhubarb
- Nuts
- Strawberries
- Raspberries
- Chocolate
- Pulses
- Beetroot

maximum of 500mg. Cholestyramine can be used as a last resort to bind intestinal oxalate, but it is often poorly tolerated.

Other factors in calcium stone formation. Although uric acid can itself form renal stones (see below), hyperuricosuria may also contribute to calcium stone formation. This may be through heterogeneous nucleation following uric acid crystallization or through other mechanisms, and this remains a controversial area. Dietary restriction of proteins and purines and/or the use of allopurinol may correct hyperuricosuria and reduce the risk of recurrence.

Citrate is a recognized inhibitor of calcium stone formation, and some stone formers have low urinary citrate concentrations as the only biochemical abnormality. This is particularly so in patients with distal renal tubular acidosis, who are already at increased risk of calcium phosphate stone formation because of failure to acidify the urine adequately. Potassium citrate supplementation is effective but should be used cautiously in patients with renal impairment or hyperkalaemia. Sodium citrate should be avoided as it can increase urinary calcium excretion.

Magnesium is also a recognized stone inhibitor. Although magnesium supplementation has not been shown to reduce stone risk, it may be worth trying if definite hypomagnesaemia is present.

Normal urine also contains a variety of protein-based inhibitors of stone formation, and it may well be that deficiencies or defects in these will be found to be the cause of some types of stone in the future. For example, a genetic variant of uromodulin (Tamm–Horsfall glycoprotein) has been found that has stone-promoting properties, whereas the normal type is thought to inhibit stone formation.

Infection-related stones

Infection-related (triple phosphate) stones are composed predominantly of magnesium ammonium phosphate (struvite), with variable amounts of calcium phosphate as carbonate–apatite. They form in the presence of high urinary concentrations of ammonia, bicarbonate and carbonate, which essentially means that they only form when the urine is infected with urea-splitting bacteria (e.g. *Proteus*, *Klebsiella* and *Pseudomonas* spp.). For this reason, and in contrast to other types of stones, they occur more frequently in women and in other people with a predisposition to urinary tract infection, such as in those with an

indwelling catheter or a spinal injury. They may also occur occasionally in other stone formers. The clinical presentation is often insidious as they tend to form staghorn calculi that may cause silent deterioration of renal function. Treatment includes correcting any underlying cause if possible, and rapid access to antibiotics appropriate to the known sensitivity of previously cultured organisms. Patients can keep a supply of antibiotics at home or even take continuous prophylaxis. Complete sterilization of the urine is possible only if stones are eradicated.

Uric acid stones

Hyperuricaemia and gout are discussed in Chapter 32. In healthy people, the amount of uric acid excreted in the urine depends on, among other things, the purine content of the diet. This makes the definition of a reference range rather difficult, particularly since even at ‘normal’ excretion rates of 3.6–4.8 mmol/24h, the urine is supersaturated with uric acid and yet most people do not form uric acid stones. Stones may form in patients with hyperuricaemia from any cause, as well as in those with renal hypouricaemia due to an isolated genetic defect or as part of a Fanconi syndrome.

Most patients with uric acid stones have normal plasma urate concentrations and urinary urate excretion, but produce concentrated urine with consistently low pH (<5.5). Patients with an ileostomy are at particularly high risk owing to bicarbonate and water loss from the stoma. The solubility of urate decreases sharply below its isoelectric point (pI) of 5.75. This offers a useful therapeutic intervention, in that if alkali is used to maintain the urine pH at 6.5–7.0, stones can be redissolved and further formation prevented even if there is a degree of hyperuricosuria. More pronounced alkalinization of the urine should be avoided, as it increases the risk of calcium phosphate deposition. Well-motivated subjects may titrate their dose of alkali (usually potassium citrate) by testing their own urine with pH paper. Patients with severe hyperuricosuria may need allopurinol to prevent stone formation and should be advised to restrict dietary purine intake.

Cystine stones

These have already been discussed in the section on cystinuria, above.

Miscellaneous rarities

Renal stones submitted for analysis occasionally may not fall into one of the categories discussed above. Some may be entirely unrelated to the renal tract (factitious disorders), whereas others may be from the renal tract but not be true stones (e.g. blood clots, sloughed papillae, encrusted sutures). However, there are a few other inherited metabolic diseases that do result in the formation of stones of unusual composition.

In hereditary xanthinuria, there is a deficiency in xanthine oxidase, with the consequent replacement of uric acid in the urine by xanthine and hypoxanthine. In about two-thirds of affected people this remains an asymptomatic metabolic abnormality, usually detected because of

very low plasma urate concentrations, but in the remaining one-third xanthine stones form in the renal tract. There may also be associated myopathy or arthritis.

An even rarer subtype of xanthinuria has been described, in which there is deficiency of xanthine oxidase together with sulphite oxidase, but here the main clinical concern is the neurological involvement.

Xanthine stones have also been described in patients with normal plasma urate concentrations who clearly do not have xanthine oxidase deficiency. The cause of these is unknown.

Another rare inherited disorder of purine metabolism associated with renal stone formation is adenine phosphoribosyl transferase (APRT) deficiency. This enzyme is involved in the salvage pathway for the purine base adenine, and deficiency results in increased urinary excretion of 2,8-dihydroxyadenine. This leads to stone formation in most homozygotes, although up to 15% remain clinically stone free. It is worth noting that many of the chemical tests used for uric acid also give a positive reaction with 2,8-dihydroxyadenine. In children, the finding of uric acid stones using chemical testing should, therefore, be further investigated using a more specific technique.

Very rarely, stones may be composed entirely of glycoprotein matrix or extraneous compounds such as triamterene, silicate or indinavir, all of which are excreted by the kidney.

Investigation of stone formers

All stone formers should have at least a minimum basic screen including serum creatinine, potassium, bicarbonate, urate and calcium with a 24 h urine collection for measurement of volume, calcium and oxalate. Evaluation should be at least six weeks after surgery or lithotripsy to avoid contamination of urine. Patients who have recurrent stones, a strong family history or who present before the age of 25 should have more extensive investigation as outlined in **Box 9.4**. The use of acid-containing bottles for urine collection is controversial because of patient safety concerns, but there is a risk of calcium oxalate and phosphate precipitation before analysis if the sample is not adequately acidified. Crystals are very difficult to redissolve, even if the urine collection is acidified on receipt in the laboratory. Accurate urate estimation requires a non-acidified collection.

There is a marked biological variation in calcium and oxalate excretion, partly owing to diet, so at least two estimations separated by a week should be requested before any treatment is started. Testing a spot urine for albumin will detect most patients with proximal tubular disorders, and can be confirmed by low molecular weight protein estimation. Urinary albumin will also be raised if there is intrinsic kidney damage or if the patient has active stone disease.

Analysis of available stones can be highly informative, although not all practitioners in the field would agree. External quality assurance surveys in the UK demonstrate that stone analysis is not generally well done, especially if the method used is qualitative. Infrared spectroscopy is the gold standard, but if this is not available, quantitative chemical techniques are adequate. Knowledge of a stone's constituent(s) obviously directs

BOX 9.4

Biochemical investigations in stone formers

Plasma

- Sodium and potassium
- Creatinine
- Bicarbonate
- Calcium and albumin
- Phosphate
- Urate
- PTH (if hypercalcaemic)

Spot urine

- pH (must be very fresh)
- Microbiology (if infection suspected)
- Amino acids
- Albumin

24 h urine*

- Volume – [>2.5 L]
- Calcium (preferably acidified) – [<6 mmol/24 h females, <7 mmol/24 h males]
- Oxalate (preferably acidified) – [<400 μ mol/24 h]
- Urate (non-acidified) – [<4 mmol/24 h]
- Citrate – [>3 mmol/24 h]
- Sodium – [<150 mmol/24 h]
- Urea – [<400 mmol/24 h]
- Magnesium – [>3 mmol/24 h]
- Creatinine – to check adequacy of collection

*Treatment target values for 24 h urine.

further biochemical investigation and monitoring. It is also important to continue to analyse the stones if there is recurrence – there may be changes in the constituents that necessitate a change in management.

Treatment

Treatment of specific stone types has been covered in the relevant sections. All patients will also benefit from general advice about diet, including those with no identifiable risk factors. Adequate fluid intake is the most important, aiming to achieve daily urine volumes of at least two but preferably three litres. Periods of dehydration, such as during sporting activities, should be avoided. Most patients can be taught to increase fluid intake to maintain a pale, straw-coloured urine. Reduction of salt, refined carbohydrate and animal protein intake has been shown to have a much greater effect on stone risk than calcium restriction. In fact, the population quartile with the lowest calcium intake has a higher risk of stone formation than the quartile with the highest intake, owing to reciprocal hyperoxaluria. Calcium restriction is therefore no longer recommended. Reducing the intake of oxalate (**Box 9.3**) will contribute to lowering of the calcium oxalate product in the urine. However, in otherwise healthy people, over 80% of urinary oxalate is produced endogenously. Tea is particularly rich in oxalate, especially if taken without milk (calcium precipitates the oxalate), and should be avoided. Citrate excretion by the kidney may be enhanced by oral potassium citrate supplements, orange juice or bicarbonate-rich mineral water. Urinary tract infections

should be treated promptly and adequately: laboratory confirmation of bacterial eradication is recommended. Long-term prophylaxis with antibiotics may be required.

CONCLUSION

Primary disorders of the renal tubules are not common, but biochemical investigations are important in both diagnosing and monitoring them.

The pathogenesis of renal stone formation is still not fully understood, but enough is known for a logical approach to be taken in the biochemical investigation of stone formers. This is something that probably has not been uniformly well done in the past.

ACKNOWLEDGEMENT

We would like to thank Stephen K. Bangert who co-authored this chapter for previous editions of the book.

Further reading

Faerber GJ. Pediatric urolithiasis. *Curr Opin Urol* 2001;11:385–9.

Detailed review of clinical research in children.

Johri N, Cooper B, Robertson W et al. An update and practical guide to renal stone management. *Nephron Clin Pract* 2010;116:c159–71.

A useful recent review with extensive references.

Pak CYC. Nephrolithiasis. *Endocrinol Metab Clin North Am* 2002;31(4).

The whole volume for this month is devoted to reviews on kidney stones, including childhood stones.

Scriver CR, Beaudet AL, Sly WS et al. editors. *The metabolic and molecular bases of inherited disease*. 8th ed New York: McGraw-Hill; 2001.

As always, the ultimate reference work for the inherited metabolic diseases. There is also a regularly updated online version available.

Unwin RJ, Capasso G. The renal tubular acidoses. *J R Soc Med* 2001;94:221–5.

An easy to understand review of the physiology, pathology, diagnosis and management of the renal tubular acidoses.

APPENDIX 9.1 DIAGNOSIS OF RENAL TUBULAR ACIDOSIS

A variety of provocative tests of urinary acidification have been used in the investigation of RTAs, but whether these can firmly diagnose the type of defect remains to be established.

In general, hyperchloraemic metabolic acidosis that is not explained by bicarbonate loss from the intestinal tract should raise the suspicion of a urinary acidification defect. The plasma potassium may give a clue as to the type (high in type 4, low in types 1 and 2). The presence of other features of the renal Fanconi syndrome suggests

type 2 RTA. The amount of bicarbonate required to correct the acidosis also gives an indication of the type of RTA (type 1 responding most readily and type 2 least readily). However, it may be necessary to confirm the diagnosis by using one of the following two tests:

Urinary acidification test

This can be used to confirm the diagnosis of distal RTA. The test is not necessary if the pH of a urine specimen collected after an overnight fast is <5.5 or if it is >5.5 in a patient confirmed to be acidotic in the resting state. If this is not the case, the patient is given ammonium chloride at a dose of 100 mg/kg body weight. Special formulations are available to minimize gastrointestinal side-effects. Urinary pH is then measured on fresh urine samples at hourly intervals for 8 h.

In normal subjects, urinary pH should fall to <5.5 in at least one sample. In distal renal tubular acidosis this does not occur and urinary pH usually remains >6.5.

Ammonium chloride should not be used in patients with liver disease, in whom calcium chloride may be used as an alternative acidifying agent (1 mmol/kg body weight). Ammonium chloride is unpleasant to take and alternative tests have been sought, for example the fludrocortisone–furosemide test. The principle of this latter test is to increase sodium delivery to the distal renal tubules at a time that sodium reabsorption and, therefore, hydrogen ion excretion is maximally stimulated by a mineralocorticoid. One mg of fludrocortisone is given orally an hour before furosemide 40 mg, also orally. Urine samples are collected every 30 min for up to 6 h or until a urine pH <5.5 is achieved. However, both false positives and false negatives can occur and ammonium chloride loading remains the definitive test.

Fractional excretion of bicarbonate

This test can be used to confirm the diagnosis of proximal RTA, as long as the patient's plasma bicarbonate is maintained above 20 mmol/L.

Plasma and urine samples are obtained and creatinine and bicarbonate concentrations measured in each. The fractional excretion of bicarbonate is then calculated:

$$\frac{\text{urine[bicarbonate]}/\text{plasma[bicarbonate]}}{\text{urine[creatinine]}/\text{plasma[creatinine]}} \times 100\%$$

In patients with proximal RTA, the fractional excretion is >10–15%, whereas in most patients with distal RTA it is <10%.

Clinical biochemistry of nutrition

Ruth M. Ayling

CHAPTER OUTLINE

INTRODUCTION 180

NUTRITIONAL REQUIREMENTS 180

The 'correct' intake 180

Energy 181

Protein 183

Micronutrients 184

Vitamins 184

Fibre 192

ASSESSMENT OF NUTRITIONAL STATUS 192

General 192

Laboratory-based assessment of individual nutrients 195

CONCLUSION 199

INTRODUCTION

Food is essential for human life. At the extremes, too little leads to starvation and too much leads to obesity, each with its associated effects on morbidity and mortality. Between these extremes, different people across the world follow apparently very different diets with no obvious resulting differences in their day-to-day health. It is apparent that in the long term the composition of the diet has an effect on the incidence of certain diseases, for example ischaemic heart disease and certain types of cancer, but often these associations are difficult to tease out because of the many confounding factors involved. However, this does not prevent very firm opinions being formulated and held about what is and what is not 'good nutrition', based on a variety of influences from basic science to religion.

Biochemistry clearly has a role in establishing the way in which the body uses various nutrients and has been important in defining certain deficiency states. Clinical biochemistry is still important in diagnosing deficiencies of certain specific nutrients, but whether there is a good biochemical marker of overall nutritional status is less clear. This chapter begins with a consideration of the various nutrients, including the effects of their deficiency or excess, and then discusses nutritional assessment. The following chapter discusses nutritional disorders and their management, both from the point of view of nutrition as an aetiological factor in disease, and disorders that are not primarily nutritional but where dietary modification or nutritional support may be important in treatment.

NUTRITIONAL REQUIREMENTS

The 'correct' intake

Part of the explanation of the fact that apparently very different diets can sustain life equally well is the concept that individual foodstuffs actually comprise different combinations of certain basic nutrients, and that it is the supply of these basic nutrients that is important rather than their origin. However, even when individual nutrients are considered, the definition of the 'correct' intake is problematical since this may be taken, for example, to be any of the following: the intake that avoids clinical signs of deficiency; the intake that maintains a given circulating concentration or tissue content of the nutrient; the intake that cures symptoms or signs of clinical deficiency; the intake that maintains a balance between intake and consumption or loss from the body over a defined period, or any one of a variety of other definitions. Even if a suitable measure can be determined for a specific nutrient, there are further problems in setting the absolute figure, since within a population, requirements will vary between individuals (owing, for example to differences in age, gender size and body composition) and even in an individual, requirements may vary in the short term (owing, for example to pregnancy, illness or environmental stress).

It is thus difficult to accurately estimate average requirements for a nutrient, and most attempts to make recommendations about intakes have depended on setting levels where most of the population will not be deficient, although it is also possible to define intakes below which most of a population will develop deficiency, and in some cases intakes that are high enough to be toxic.

TABLE 10.1 Definitions of terms used in describing nutrient requirements

Abbreviation	Term	Definition
DRV	Dietary reference value	General term that includes all those listed and some more pragmatic approaches
RDI RDA	Recommended daily intake Recommended daily amount	Average intakes that should meet the needs of almost all members of a population. In the UK, these are based on the Department of Health and Social Security publications: <i>Recommended intakes of nutrients for the UK and Recommended daily amounts of food energy and nutrients for groups of people in the UK</i>
EAR	Estimated average requirement	Mean requirement for a population: about half will need more and half less
RNI LRNI	Reference nutrient intake Lower reference nutrient intake	Intakes two standard deviations above (RNI) and below (LRNI) the EAR for a population. Assuming individual requirements are normally distributed, these limits should define requirements for 95% of the population

Adapted from Department of Health 1991.

These points are covered in more detail in some of the texts listed under Further reading, below, and no attempt will be made here to define ideal intakes in absolute terms. Some of the terms used to describe nutrient requirements are defined in Table 10.1.

Energy

The body uses energy in the maintenance of metabolic processes, in physical activity and in growth. In the resting individual, energy-requiring processes include the active pumping of ions across cell membranes, thermoregulation, cell division and basal function of all of the body's systems. Resting energy expenditure appears to alter with energy intake to some extent. Physical activity, particularly if vigorous, can increase energy expenditure markedly, although such activity is not usually sustained for long enough to make the increase substantial when measured over a 24 h period. The laying down of new tissues represents an investment of energy, so it is not surprising that energy requirements, expressed in terms of body weight, are highest in infants and young children, with a gradual decline from the third decade into old age.

Thus, the energy requirement of an individual varies with body size and composition, gender, age, nutritional status and climate. In women, energy expenditure is increased during pregnancy and lactation, and in any individual, illness or response to trauma (e.g. systemic infection, burns) may cause a considerable increase.

Energy expenditure and energy intake do not always rise and fall in parallel. A sustained deficiency in energy intake generally leads to consumption of body stores of energy, including protein as well as glycogen and fat. Excessive energy intake (if sustained) results in obesity. Both of these conditions are considered further in the next chapter.

Dietary energy is mainly obtained from a combination of fat and carbohydrate. In a Western diet, each may account for 40–50% of total energy intake, although in developing countries the proportion from carbohydrate tends to be higher and that from fat, lower. Although dietary protein is generally thought of as a source of

materials (amino acids) for endogenous protein production, it may also be used for energy production, for example if protein intake is adequate but non-protein energy intake is low. Ethanol also has to be considered as a significant energy source in certain populations; for example, in the UK, it forms 7% of average energy intake. (Interestingly, at higher ethanol intakes – 25–35% of dietary energy – ethanol appears not to be completely utilized as a source of energy, although the reason for this is not clear.)

Carbohydrate

Dietary carbohydrate provides 4.1 kcal/g and most is derived from sugars and starch. The sugars are mainly the monosaccharides glucose, fructose and galactose, and the disaccharides sucrose, lactose and maltose. Dietary sugars can be divided into those which are present in intact cells, such as in whole fruit (intrinsic sugars), and those which are free and readily absorbed as a result of having been added to food, usually as sucrose (extrinsic sugars). Milk sugars are usually considered as intrinsic. A high intake of extrinsic sugars is associated with an increased prevalence of dental caries. Sugar derivatives, such as the sugar alcohols sorbitol and xylitol, can be partially digested and can provide 2.4 kcal/g.

In the UK, it is recommended that in the adult, 40–50% of energy intake should be provided by carbohydrate, but non-milk extrinsic sugars should not exceed 11% of energy intake. There is considerable debate about whether a high intake of non-milk extrinsic sugars may have harmful effects in addition to the increased risk of dental caries. If such an intake results in an increased energy intake, then obesity is likely to be a problem, but in any case, ingestion of large amounts of extrinsic sugars can lead to raised plasma glucose, insulin and lipid concentrations, all of which are potentially harmful.

Large amounts of starch are not harmful: indeed, in developing countries, starch may form 75–80% of the total energy intake. Starches are α -glucan polysaccharides, of which there are two major forms, amylose and amylopectin (see Fig. 10.1). They are partially resistant to

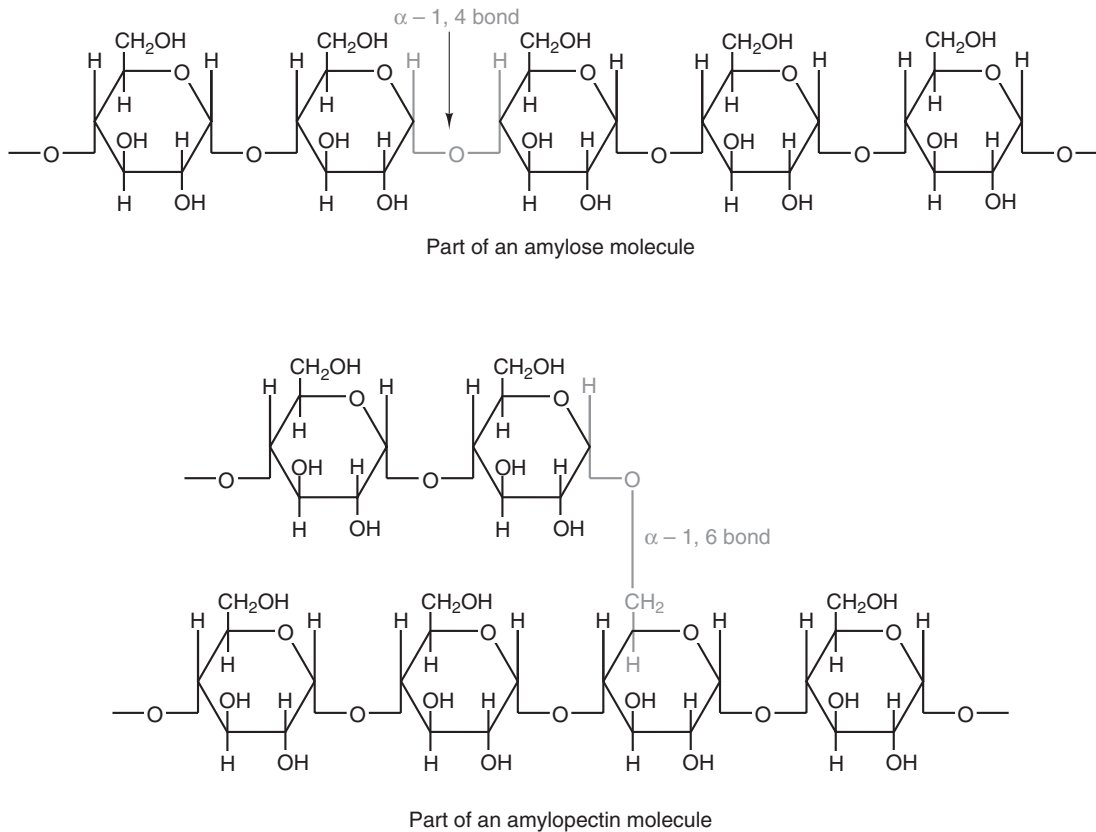


FIGURE 10.1 ■ Molecular structure of starches. Amylose is a straight chain glucose polymer with a molecular weight of 100 000 Da or more and amylopectin is a branched chain glucose polymer with a molecular weight of over 1 million Da.

hydrolysis during digestion and are thus not totally available as an energy source. Certain other carbohydrates, that are not energy sources, but form part of 'dietary fibre', are considered later.

A deficiency in dietary carbohydrate will either lead to energy deficiency or result in potentially harmful amounts of other energy sources being required in the diet if energy intake is to be maintained. However, certain tissues have an obligatory requirement for glucose as an energy substrate (e.g. brain and nervous system (in the short term), red blood cells) and, while this need can be met by gluconeogenesis, some carbohydrate is necessary in the diet if ketosis is to be avoided.

Fat

The term 'fat' includes a range of substances, including triacylglycerols (triglycerides), phospholipids and sterols (e.g. cholesterol). Triacylglycerols are the most common storage fats and are thus the most important fatty energy source in foods. They are a relatively 'concentrated' energy source, with a higher energy content per unit weight (9.3 kcal/g) than carbohydrate (4.1 kcal/g). Although all triacylglycerols can act in this way, there may be important long-term health differences between them, dependent on the structure of the fatty acid residues they contain. The main differences are in the length of the

carbon chain, the presence (and number) of any unsaturated bonds and, if such bonds are present, the positional and geometric isomers present. Some examples of these differences are illustrated in Figure 10.2. The effects of excessive intake of triglycerides are discussed further in Chapter 11, and the differences in absorption and metabolism between long chain and medium chain triglycerides in Chapter 12.

Dietary fat deficiency does not generally seem to be a problem, as most fats necessary to the body can be endogenously synthesized when a high proportion of total energy intake is supplied by carbohydrate. However, there are certain fatty acids that appear to be essential, at least in small amounts. These are linoleic acid (C18:2, ω -6) (for nomenclature, see Fig. 10.2) and α -linolenic acid (C18:3, ω -3). They are important components of phospholipids, in which they help to maintain the function of cellular and subcellular membranes. They are also involved in the regulation of cholesterol transport, breakdown and excretion. They are the precursors of arachidonic (C20:4, ω -6), eicosapentaenoic (C20:5, ω -3) and docosahexaenoic (C22:6, ω -3) acids and are thus also important in the synthesis of prostaglandins, thromboxanes and leukotrienes. These three longer chain fatty acids are not strictly essential fatty acids (EFAs), but may become conditionally so in EFA deficiency. Adequate dietary supply of the longer chain fatty acids is probably also important during rapid

brain growth in infancy, particularly in fast-growing premature infants.

The clinical effects of EFA deficiency include dermatitis, alopecia and fatty liver, but these are only seen when EFAs provide less than 1–2% of total dietary energy intake. In patients receiving long-term fat-free parenteral nutrition, cutaneous application of small amounts of appropriate oils has alleviated biochemical EFA deficiency.

SATURATED FATTY ACIDS

$\text{CH}_3 (\text{CH}_2)_{14} \text{COOH}$
Palmitic (hexadecanoic) acid (C16:0)

$\text{CH}_3 (\text{CH}_2)_{16} \text{COOH}$
Stearic (octadecanoic) acid (C18:0)

MONOUNSATURATED FATTY ACIDS

$\text{CH}_3 (\text{CH}_2)_5 \text{CH}=\text{CH}(\text{CH}_2)_7 \text{COOH}$
Palmitoleic (Δ^9 hexadecenoic) acid (C16:1)

$\text{CH}_3 (\text{CH}_2)_7 \text{CH}=\text{CH}(\text{CH}_2)_7 \text{COOH}$
Oleic (Δ^9 octadecenoic) acid (C18:1)

POLYUNSATURATED FATTY ACIDS

ω -6
 $\text{CH}_3 (\text{CH}_2)_4 \text{CH}=\text{CHCH}_2 \text{CH}=\text{CH}(\text{CH}_2)_7 \text{COOH}$
Linoleic ($\Delta^{9,12}$ octadecadienoic) acid (C18:2)

ω -3
 $\text{CH}_3 \text{CH}_2 \text{CH}=\text{CHCH}_2 \text{CH}=\text{CHCH}_2 \text{CH}=\text{CH}(\text{CH}_2)_7 \text{COOH}$
 α linoleic ($\Delta^{9,12,15}$ octadecatrienoic) acid (C18:3)

ω -6
 $\text{CH}_3 (\text{CH}_2)_4 \text{CH}=\text{CHCH}_2 \text{CH}=\text{CHCH}_2 \text{CH}=\text{CHCH}_2 \text{CH}=\text{CH}(\text{CH}_2)_3 \text{COOH}$
Arachidonic ($\Delta^{5,8,11,14}$ eicosatetraenoic) acid (C20:4)

FIGURE 10.2 ■ Molecular structure of some fatty acids. The systematic name (in brackets) indicates the number of carbon atoms in the chain (e.g. octadeca- = 18) and the presence and number of any double bonds (e.g. -anoic = 0, -enoic = 1, -dienoic = 2, -trienoic = 3 etc.). This is sometimes abbreviated to (Cx:y), x and y indicating the numbers of carbon atoms and double bonds respectively. The numbers following the Δ symbol in the systematic names indicate the positions of the double bonds, for example, Δ^9 indicates a double bond between carbons 9 and 10, counting the carboxylic acid carbon as 1. An alternative system counts from the methyl carbon to the first double bond (e.g. ω -3). Most double bonds in naturally occurring unsaturated fatty acids are in the *cis* configuration, although industrial processing of vegetable oils to produce saturated fatty acids (hydrogenation) can also cause isomerization to the *trans* position.

Protein

Proteins have a key structural and functional role in virtually all bodily processes, and a supply of appropriate amino acids is necessary during normal turnover, with additional amounts if extra protein is being formed, for example during growth, pregnancy or lactation. This supply generally comes from dietary protein, of either animal or vegetable origin, but the actual amount of protein required depends on its amino acid content. Human protein contains 20 amino acids (Box 10.1), of which nine are recognized to be essential (or ‘indispensable’) in adults – the remaining 11 can be synthesized endogenously if the food protein is deficient in these but contains sufficient essential amino acids. Some amino acids, for example cysteine and glutamine, whilst not

BOX 10.1

Essential and non-essential α -amino acids found in human proteins

Designations can change under certain circumstances, for example phenylalanine and methionine are necessary for the synthesis of tyrosine and cysteine, respectively; if either phenylalanine or methionine is in short supply, then each of the other two can become ‘conditionally’ essential. Other amino acids (e.g. ornithine, citrulline, taurine) occur in the body, but are not found in proteins.

Essential amino acids

Histidine
Isoleucine
Leucine
Lysine
Methionine
Phenylalanine
Threonine
Tryptophan
Valine

Non-essential amino acids

Alanine
Arginine
Asparagine
Aspartic acid
Cysteine
Glutamic acid
Glutamine
Glycine
Proline^a
Serine
Tyrosine

^aProline is actually an imino acid but by convention, is included here.

normally essential, may become so in conditions such as severe illness, trauma or burns, when requirements exceed the body’s ability to synthesize them. In children, due to the increased demands of growth, the threshold for an amino acid to become conditionally essential is lower. Only L-amino acids are incorporated into human proteins.

Some proteins, particularly those derived from animal sources such as eggs, milk, fish and meat, contain all the essential amino acids in sufficient amounts for protein synthesis and are said to be complete, or of high quality. Foods lacking one or more amino acids are termed incomplete proteins; these are often from plant sources such as pulses. In diets in which plant foods are the main source of protein, combining different types of food in appropriate quantities is required in order to obtain all the essential amino acids. However, the amino acid profile of dietary proteins does not completely predict the amounts of these amino acids that will be available for protein synthesis after digestion and absorption: some of them may be biologically unavailable. Lysine is particularly important in this regard, as its supply in many dietary proteins is relatively low in comparison to nutritional requirements. It becomes unavailable if its ϵ -amino group reacts with another molecule, because ϵ -amino bonds are not cleaved by digestive enzymes. Such bonds may form with other amino acids (e.g. the carboxyl side chains of glutamic or aspartic acids) or with reducing sugars (during cooking or storage of foods).

Deficiency of protein generally occurs as part of more generalized malnutrition and is discussed later. Excessive protein intake in renal disease accelerates deterioration in renal function (see Chapter 7) and it is possible that it may have effects in healthy people, for example on renal function and on bone mineral density.

Micronutrients

Vitamins and certain trace elements are essential components of the diet required in very small amounts. Deficiency of individual micronutrients classically results in typical symptoms and signs according to the vitamin or trace element involved; recognition and treatment of deficiency of a single micronutrient deficiency is relatively easy. However, milder forms of deficiency, particularly of multiple micronutrients concurrently, may be difficult to recognize clinically.

Micronutrient status is likely to be affected by acute illness, owing to a combination of reduced intake and increased demand. Such increased demand may be due to catabolism resulting from the illness, subsequent anabolism and from disease-specific increased losses, for example loss from fistulae, diarrhoea and dialysis. As laboratory concentration of micronutrients may be perturbed in the acute phase, deficiencies in this group of patients can be particularly difficult to assess.

It is increasingly being recognized that subclinical deficiencies of micronutrients may have detrimental effects, some of which are described in the following pages. As the assessment of micronutrients can be important in patients receiving supplementation, consideration is also given to aspects of micronutrient toxicity.

Vitamin D is considered with bone metabolism in Chapter 31. Cobalt as a constituent of vitamin B₁₂ is discussed, together with this vitamin, folate and iron, in Chapter 27. Iodine is essential in the formation of thyroid hormones (see Chapter 19).

Vitamins

Fat-soluble vitamins

Vitamins are traditionally classified into those that are fat soluble (A, D, E and K) and the remainder, which are

water soluble. This is probably still a useful distinction, as it predicts when deficiency is likely (e.g. fat-soluble vitamins in steatorrhoea), and is retained here.

The fat-soluble vitamins A, E and K are structurally different, but are all non-polar, water-insoluble lipids. Absorption of each of them requires micelle formation in the gut lumen, for which bile salts are necessary.

Vitamin A. Dietary vitamin A occurs in a variety of forms which can be broadly divided into preformed vitamin A (retinol) and carotenoids, which are cleaved in the body to give vitamin A (see Fig. 10.3). In the UK, up to 75% of dietary vitamin A is preformed, derived from foods such as fortified cereals, margarine and dairy products, fish-liver oil and multivitamin preparations. Carotenoids are found in plants; the most important in the UK diet is β -carotene, which is found in carrots, dark green leafy vegetables, pumpkins and mangoes.

Retinol is transported in the bloodstream bound to retinol-binding protein (RBP), which in turn forms a complex with thyroid-binding pre-albumin. Protein-energy malnutrition can result in reduced synthesis of RBP and thus impaired retinol transport from the liver, producing a functional vitamin A deficiency, even in the presence of adequate liver reserves. β -Carotene undergoes oxidative fission in the intestine to give retinal (see Fig. 10.3), and hence retinol. The cleavage of provitamin A carotenoids appears to be inhibited when vitamin A stores are high, hence, toxicity from ingestion of plant sources is rare. Ingestion of excess preformed vitamin A leads to its absorption and hepatic storage.

The best understood of the actions of vitamin A is probably its role in vision. The 11-*cis* form of retinal is a component of the visual pigment rhodopsin, which is found in the rods of the retina. Light causes the retinal to change to the all-*trans* form (see Fig. 10.3) and this triggers a series of conformational and other changes, the

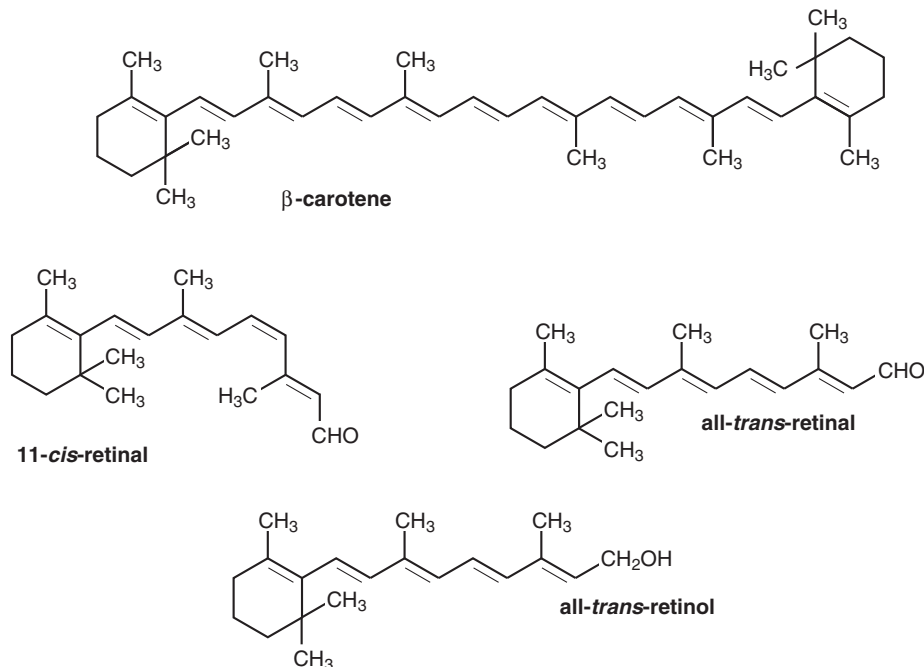


FIGURE 10.3 ■ Structure of vitamin A and some related compounds.

end result being that an electrical signal is transmitted to the cortex of the brain and is perceived as light. A similar mechanism operates in the cones of the retina.

Vitamin A is also necessary for growth and for normal development and differentiation of tissues. These actions are mediated by retinoic acid, which is present in foods in only small amounts but can be manufactured in the body from retinol. It modulates gene expression by activation of nuclear receptors, of which there are two groups: retinoic acid receptors (RAR) bind all-*trans*-retinoic acid (and some other retinoids); retinoid X receptors (RXR) bind 9-*cis*-retinoic acid. The latter interact with other receptors (e.g. those for vitamin D and thyroid hormones), and in the absence of retinoic acid can also act as repressors of gene expression. Other metabolic roles of vitamin A include carrying mannosyl units in the synthesis of hydrophobic glycoproteins and the retinoylation of proteins, for example the regulatory subunits of cAMP protein kinases.

Vitamin A deficiency is common in some developing countries, particularly in children, and is the single most common preventable cause of blindness in the world. An early symptom is night blindness, because of inadequate amounts of 11-*cis*-retinal in the retina. However, deficiency also has important effects in mucus-secreting epithelia and so there are other consequences for the eyes. Keratinizing squamous metaplasia in the conjunctiva causes conjunctival xerosis and this may be followed by the appearance of white plaques of desquamated thickened epithelium, known as Bitot's spots. Similar metaplasia in the cornea can lead to ulceration, which may progress to scarring and consequent blindness. Changes in other mucus-secreting epithelia, for example in the respiratory and gastrointestinal tracts, probably account in part for the lowered resistance to infections reported in vitamin A deficiency. In such areas of the world, public health programmes aimed at administering large doses of vitamin A have been successful in reducing morbidity and mortality associated with its deficiency, particularly in young children.

Acute ingestion of large amounts of preformed vitamin A can result in raised intracranial pressure, with headaches, nausea, vomiting and visual disturbances. Chronic overdosage is associated with bone damage including increased bone resorption, spontaneous fracture and sometimes raised plasma calcium concentration and alkaline phosphatase activity. Liver damage, hair loss and skin changes have also been reported.

Retinol is teratogenic (or, at least, the synthetic retinoids used in the treatment of certain dermatological conditions are) and so, in the UK, pregnant women are advised against self-medication with vitamin A and against

consumption of liver and products made from it. A high intake of β -carotene results in an orange-yellow appearance of the plasma, body fat and skin, but is not usually regarded as harmful in the short term. However, there may be an increased incidence of malignancy in people taking β -carotene supplements, particularly among heavy smokers.

Vitamin E. There are eight very similar compounds that have vitamin E activity, four tocopherols and four tocotrienols. The most active is the natural isomer of α -tocopherol, which accounts for about 90% of the vitamin E present in human tissues. Its structure is shown in Figure 10.4. It appears to be the major lipid-soluble antioxidant in cell membranes, acting to prevent the peroxidation of unsaturated fatty acids by free oxygen radicals. α -Tocopherol, but not its related tocopherols and tocotrienols, also has actions that are independent of its antioxidant properties. It modulates the transcription of certain genes, inhibits platelet aggregation and vascular smooth muscle proliferation and has an effect on cell signalling in the immune system.

The nutritional requirement for vitamin E is approximately proportional to the intake of polyunsaturated fatty acids. However, since foods rich in these (e.g. vegetable oils) also tend to contain large amounts of vitamin E, deficiency states are rare.

Vitamin E is not easily transported across the placenta and the first vitamin E deficiency state to be firmly established was in premature infants, who developed haemolytic anaemia, thrombocytosis and oedema. In children and adults who are unable to absorb or utilize vitamin E adequately (e.g. in cystic fibrosis or abetalipoproteinaemia), a progressive spinocerebellar degeneration may develop. The full syndrome consists of ataxia of the limbs, loss of position and vibration senses, absent deep tendon reflexes and pigmentary degeneration of the retina. Treatment with vitamin E supplements can prevent these neurological features or, if they are already present, arrest or even reverse them. For α -tocopherol to be made available to the tissues from very low density lipoprotein, it must first bind to hepatic α -tocopherol transfer protein. Individuals with a deficiency of this transfer protein develop a very similar clinical syndrome (ataxia with vitamin E deficiency,AVED).

There is some evidence that increased tissue concentrations of antioxidants, in particular vitamin E, may protect against conditions such as cancer and ischaemic heart disease, but this remains controversial.

Vitamin E supplements have also been taken to improve general well-being and sexual performance, but the only firm conclusion that can be drawn is that quite high

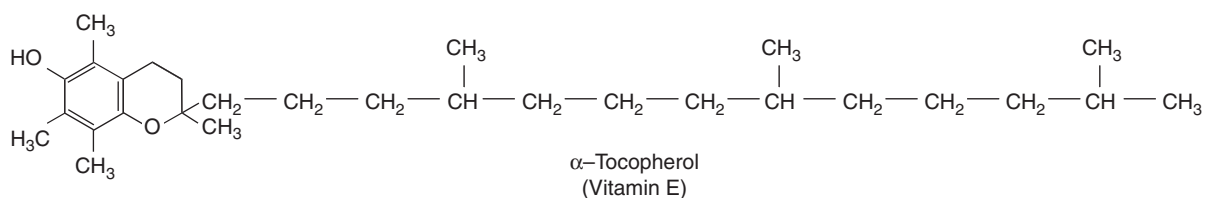


FIGURE 10.4 ■ Structure of α -tocopherol. It is an alcohol of empirical formula $C_{29}H_{50}O_2$, with a long alkyl tail, which confers fat solubility. The hydroxyl group gives it its characteristic antioxidant property.

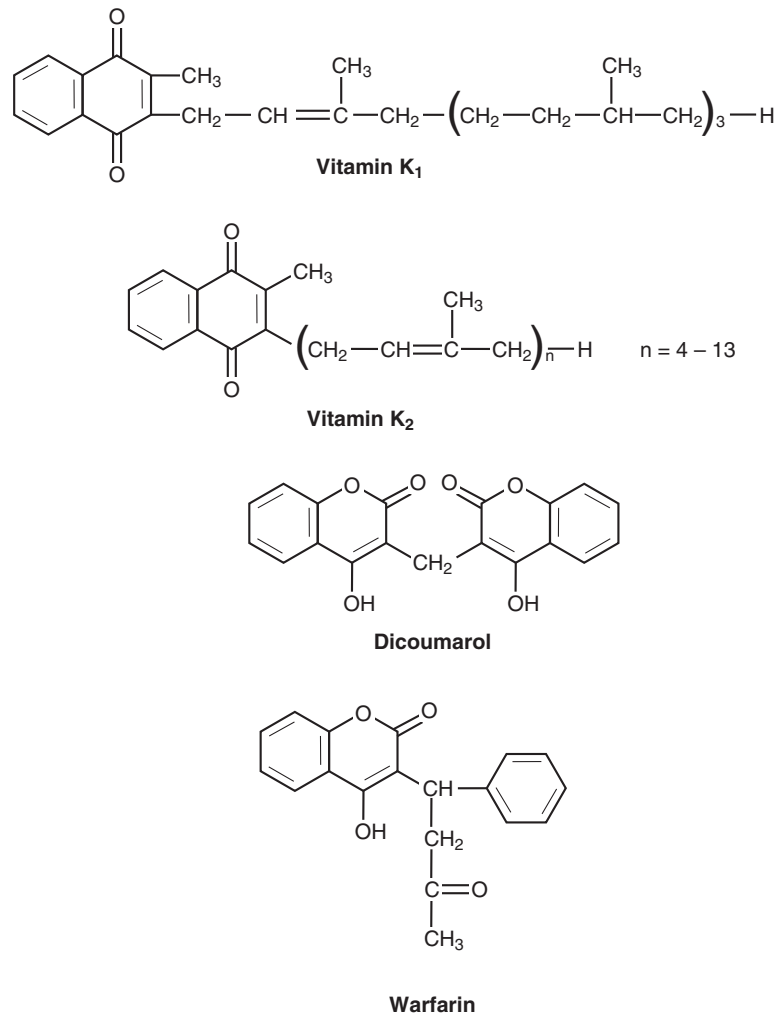


FIGURE 10.5 ■ Structure of vitamin K₁ (phylomenadione) and vitamin K₂ (menaquinone), together with two vitamin K antagonists that have anticoagulant properties.

intakes of vitamin E are non-toxic, although very high intakes may antagonize vitamin K and potentiate anticoagulant therapy.

Vitamin K. The structures of the two forms of vitamin K and two vitamin K antagonists are shown in [Figure 10.5](#). Vitamin K is involved in the post-translational modification of the blood coagulation factors II (prothrombin), VII, IX and X. A vitamin K-dependent enzyme system catalyses the carboxylation of the first ten amino-terminal glutamate residues in prothrombin to γ -carboxyglutamate (Gla). This carboxylation facilitates the binding of calcium, which is necessary for prothrombin to bind to platelet phospholipid, which is in turn necessary for the conversion of prothrombin to thrombin in the final common pathway of the clotting cascade. A similar process of γ -carboxylation occurs in factors VII, IX and X to form sites with a high affinity for calcium, and this is also known to occur in some clotting inhibitors (protein C and protein S). It also occurs in some proteins outside the clotting system, for example in osteocalcin and the matrix Gla protein in bone, in nephrocalcin in the renal cortex and in the product of the growth arrest specific

gene (*GAS6*), which regulates differentiation and development in nervous tissue and apoptosis in other tissues.

The main dietary form of vitamin K is phylomenadione (K₁), with leafy green vegetables being the richest source, although other vegetable sources are important in most diets. Menaquinones (K₂) produced by the intestinal bacteria may also be important, since while circulating vitamin K is mainly phylomenadione, hepatic reserves are mainly menaquinones.

Vitamin K deficiency in its severest form leads to a bleeding syndrome, but dietary deficiency is rare after the first few months of life, unless there is an underlying disease affecting the absorption or utilization of the vitamin. However, vitamin K is undetectable in the blood of new-born babies, their hepatic reserves are low compared with adults and human breast milk contains barely adequate amounts. Thus, deficiency may present in infants as vitamin K deficiency bleeding (VKDB), within the first 24h from the time of birth (early), between two and seven days (classic) or after seven days (late). Intracranial haemorrhage is rare in the early form, but occurs in more than 50% of babies with the late form and is likely to lead to death or severe neurological

sequelae. The administration of vitamin K in the neonatal period abolishes the risk of deficiency. A single intramuscular injection of 1 mg of phytomenadione is highly effective for this purpose but suspicions, which now appear to be unfounded, were raised that its use led to an increased risk of childhood malignancy, particularly leukaemia. This adverse association has not been found for the administration of oral vitamin K, but the latter requires a multi-dose regimen. The current UK recommendation is that either intramuscular or oral vitamin K should be offered to all neonates after informed discussion with the parents.

Toxicity from naturally occurring K vitamins would appear to be rare, even in quite high doses. However, the use of synthetic preparations of menadione for nutritional purposes is best avoided, as serious unwanted effects have been recorded (e.g. haemolysis and liver damage in neonates).

Water-soluble vitamins

The water-soluble vitamins are polar compounds that, with the exception of vitamin B₁₂, tend to be absorbed rapidly from the upper small intestine, mainly through either sodium-dependent active transport or carrier-mediated diffusion. Vitamin B₁₂ and folate are discussed in Chapter 26.

Thiamin. The physiologically active form of thiamin (vitamin B₁) is thiamin pyrophosphate (TPP). This functions as a cofactor in the conversion of pyruvate to acetyl-CoA by the pyruvate dehydrogenase complex and in the conversion of 2-oxoglutarate to succinyl-CoA by the 2-oxoglutarate dehydrogenase complex. It acts as a cofactor to the enzyme transketolase in the pentose phosphate pathway and plays a part in the metabolism of branched chain amino acids. Thiamin also has non-cofactor roles: thiamin triphosphate is found in nerve and muscle cells, where it activates membrane ion channels, possibly by phosphorylating them.

Thiamin is found in most foodstuffs, but wheatgerm, oatmeal and yeast are particularly rich sources. However, the body store of about 30 mg is only 30 times the daily requirement and so an inadequate diet is likely to lead to deficiency of thiamin sooner than of any other vitamin. Thiamin deficiency used to be widespread in areas where rice formed a major part of the diet, resulting from the consumption of polished rice (from which the husk had been removed). In present times, thiamin deficiency is seen more typically in alcoholics (alcohol decreases the absorption of thiamin). Diets high in carbohydrate require more thiamin for their assimilation than diets high in fat, and subclinical deficiency may be unmasked by refeeding with a carbohydrate-rich diet. Patients on long-term renal replacement treatment may become deficient in thiamin (and other water-soluble vitamins) unless given supplements.

Most of the clinical features of the disease beriberi respond to treatment with thiamin, suggesting that it is mainly due to thiamin deficiency. Two forms are described, one in which there is peripheral neuropathy, muscle weakness, general fatigue and reduced

attention span ('dry' beriberi) and one in which there is also oedema and heart failure ('wet' beriberi). Some of these features may be the result of coexisting protein deficiency.

Patients with thiamin deficiency (commonly alcoholics) may present with a neuropathy and cardiomyopathy, but may also develop an encephalopathy, the Wernicke–Korsakoff syndrome. Wernicke encephalopathy and Korsakoff psychosis were originally described as two separate conditions, but now appear to be the acute and chronic manifestations, respectively, of a single condition. Wernicke encephalopathy has an acute onset, with ophthalmoplegia, nystagmus, ataxia and stupor or apathy. It is a medical emergency: the ophthalmoplegia, ataxia and lowered consciousness respond, in most cases within two days, to treatment with thiamin but, in up to 80% of cases, the mental changes fail to resolve completely and Korsakoff psychosis develops. Once established, this responds only slowly or not at all to thiamin treatment. The main features are retrograde amnesia, difficulty in assimilating new ideas and confabulation. Susceptibility to the development of this syndrome in thiamin deficiency may be greater in people who have a genetic variant of transketolase that binds thiamin less avidly than usual.

Thiamin is relatively non-toxic and can be given safely if deficiency is suspected. However, anaphylaxis has occurred after parenteral administration, and a wide variety of toxic effects has been described in adults with chronic intakes in excess of 50 mg/kg body weight or 3 g/24 h.

Riboflavin. Riboflavin (Vitamin B₂) is a constituent of the two flavins found in flavoproteins, flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD), each of which acts as an electron carrier in many vital biological oxidoreduction systems. It is also a constituent of the cryptochromes, blue-light sensitive pigments in the eye that are important in the setting and maintenance of circadian rhythms. Clinical riboflavin deficiency is rare in countries where milk is a regular part of the diet; of-fal is another rich source. Even when dietary deficiency is present, clinical effects are rare because any riboflavin present in the body can be very efficiently conserved and reutilized.

The clinical features of riboflavin deficiency recorded in volunteers on a riboflavin-deficient diet include angular stomatitis, cheilosis, atrophic lingual papillae, glossitis, magenta tongue, seborrhoeic skin lesions, superficial lesions of the genitalia and vascularization of the cornea. The biochemical basis of these clinical manifestations is not immediately obvious and it may be that riboflavin deficiency has effects on the metabolism of other nutrients, for example pyridoxine, iron and folate.

Riboflavin absorption from the gastrointestinal tract is readily saturated, and any excess that does occur is rapidly excreted in the urine, so riboflavin does not accumulate in the body even when the oral intake is very high. This is probably just as well, since as well as reoxidation of reduced flavin coenzymes being a major source of oxygen radicals in the body, riboflavin is itself capable of non-enzymically generating reactive oxygen species.

Nicotinamide. Nicotinamide is a constituent of nicotinamide adenine dinucleotide (NAD⁺) and its phosphate (NADP⁺).

These are involved in a large number of oxidoreduction reactions in both cytosol and mitochondria, undergoing reversible reduction to NADH and NADPH, respectively. The major role of NADH is to transfer electrons from metabolic intermediates into the electron transfer chain, while NADPH acts as a reducing agent in a large number of biosynthetic processes. NAD⁺ and NADP⁺ are commonly referred to as coenzymes, but are probably better considered as true substrates.

Nicotinamide and its precursor nicotinic acid are both plentiful in animal and plant foods, although in some plant sources, for example maize, they are in a bound form and biologically unavailable. In man, a small amount is also formed in a minor pathway of tryptophan catabolism, via kynurenine to nicotinate, and it is usually accepted that 60 mg of tryptophan is equivalent to 1 mg of dietary nicotinamide or nicotinic acid. Both dietary and endogenous sources appear to be relatively important, since the deficiency state, pellagra, may occur when either is decreased. Simple dietary deficiency is now rare unless maize forms the main dietary constituent, but may be a complication of therapeutic low-protein diets. Decreased formation from tryptophan may occur where there is coexisting pyridoxine or riboflavin deficiency, since kynureninase and kynurenine hydroxylase are both dependent on these. Oestrogens also decrease the rate of tryptophan metabolism, so premenopausal women are more susceptible than men to borderline dietary deficiency. In Hartnup disease, there is decreased tryptophan absorption from the gut, and in carcinoid syndrome, there is increased usage of tryptophan in the synthesis of large amounts of 5-hydroxytryptamine – patients with either of these conditions are, therefore, particularly susceptible to decreases in nicotinamide and nicotinic acid intake.

The initial features of pellagra are non-specific (weakness, lassitude, anorexia and indigestion), but are followed by pigmented dermatitis in areas of skin exposed to sunlight, diarrhoea with widespread inflammation of epithelial surfaces and dementia, which may be preceded by irritability and depression.

Nicotinic acid is sometimes used in high doses for the treatment of combined hyperlipidaemia and the adverse effects are known to include flushing, hepatotoxicity, hyperuricaemia and impaired glucose tolerance.

'Niacin' is sometimes used to mean both nicotinic acid and nicotinamide and sometimes just one or other of them; because of this imprecision, the term has not been used here.

Vitamin B₆. Pyridoxal, pyridoxine, pyridoxamine and their 5'-phosphates (Fig. 10.6) are interconvertible in the body and all have vitamin B₆ activity. Pyridoxal phosphate is a cofactor for over 60 enzymes that catalyse reactions of amino acids, and so the absolute requirement is related to the rate of amino acid metabolism. It is also a cofactor for glycogen phosphorylase, and has roles in the modulation of steroid hormone action and the regulation of gene expression. Primary dietary deficiency of vitamin B₆ is rare, since it is widely distributed in foods (although in some vegetable sources, it may be present in a biologically unavailable glycoside form), body stores of the vitamin are reasonable and it is synthesized by intestinal flora. The first definite cases of deficiency described were in infants fed a milk preparation that had been overheated during manufacture, destroying much of the vitamin B₆. Some (but not all) of the infants developed neurological symptoms, including convulsions, which responded to vitamin B₆ supplements. Human volunteers and animals fed a B₆-deficient diet develop neurological symptoms together with changes in the mouth and skin. The changes in the nervous system are presumably due to the role of vitamin B₆ in the metabolism of neurotransmitters. Certain drugs are known to interfere with vitamin B₆ metabolism, for example isoniazid, which binds to pyridoxal-5-phosphate reducing its availability, and penicillamine; vitamin B₆ supplementation during drug treatment is advised.

Pharmacological doses of pyridoxine may be useful in certain rare inherited conditions, for example hypochromic sideroblastic anaemia, but in others (e.g. premenstrual tension), the evidence for benefit is less good, and in very high doses, pyridoxine may itself cause a severe peripheral sensory neuropathy.

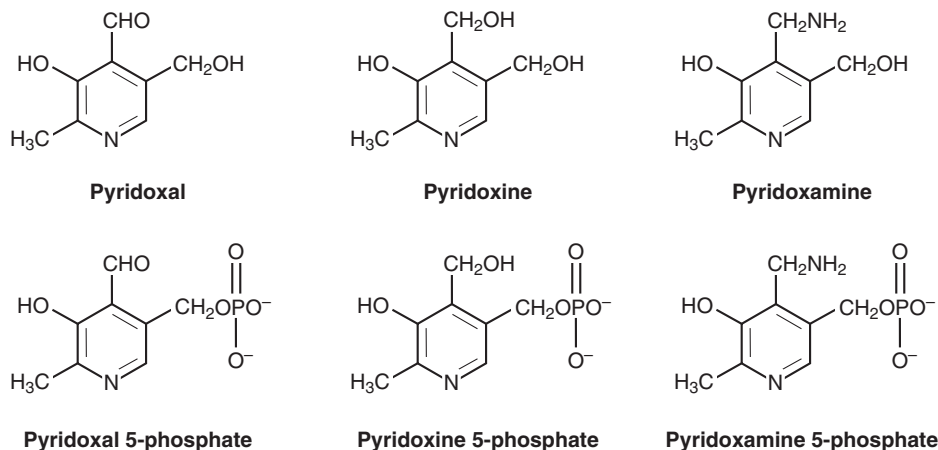


FIGURE 10.6 ■ Structures of various forms of vitamin B₆.

Pantothenic acid. Pantothenic acid is a constituent of the coenzyme A molecule and is thus involved in many major metabolic pathways. It is also present (not as part of coenzyme A) in fatty acid synthase.

The name means 'available everywhere' and this reflects its distribution in foods. Particularly rich sources include liver, meat, cereals, milk, egg yolk and fresh vegetables. Spontaneous isolated deficiency in man has never been proven, but experimental deficiency in human volunteers leads to vomiting, malaise, abdominal distress and burning cramps, and later on, tenderness in the heels, fatigue and insomnia.

Biotin. Biotin acts as a coenzyme in a number of one-carbon transfer reactions. ATP hydrolysis is coupled to the carboxylation of the biotin to form N-1'-carboxybiotin, which then carboxylates the substrate of the enzyme. Examples include pyruvate carboxylase and acetyl-CoA carboxylase, so biotin is important in gluconeogenesis and lipogenesis and also plays a part in the catabolism of branched chain amino acids.

Biotin is widely distributed in foods, with offal, milk and eggs being particularly rich sources. Intestinal bacteria synthesize biotin, but it is not clear how much of this is available for absorption. Deficiency has been described in patients on parenteral nutrition deficient in biotin and also in people who consume large amounts of raw eggs: egg white contains a glycoprotein called avidin that binds biotin and prevents absorption. Such patients develop a fine scaly dermatitis and hair loss, but isolated biotin deficiency in man is otherwise unknown, apart from a suggestion that subclinical deficiency in pregnant women may be an aetiological factor in some birth defects. Even large doses of biotin appear to be non-toxic.

Vitamin C. Vitamin C comprises L-ascorbic acid and the oxidized derivative, L-dehydroascorbic acid. In animal tissues, 90% is in the form of ascorbic acid, but the two are interconvertible and both are biologically active. Man is unusual in that most other mammals can synthesize ascorbate from D-glucose or D-galactose. Guinea pigs cannot and so form a useful experimental model.

Vitamin C is an important aqueous antioxidant in the body, although its relationships with other antioxidants remain to be fully elucidated. Its best characterized role is in the post-translational reduction of proline to form hydroxyproline. This imino acid is relatively uncommon in the body apart from in collagen, where it is essential, and one of the effects of vitamin C deficiency is impaired collagen formation. Vitamin C also assists in the intestinal absorption of non-haem iron, by keeping it in the ferrous (Fe^{2+}) form.

The main dietary sources of vitamin C are citrus and soft fruits, together with the growing points of certain vegetables, although, in the UK, potatoes also make a significant contribution. It is one of the most labile of nutrients, being destroyed by oxygen, metal ions, alkaline conditions, heat and light: cooking reduces the amount of dietary ascorbic acid available for absorption.

Deficiency of vitamin C causes scurvy, with perifollicular (or more extensive) haemorrhages, bleeding gums, poor wound healing, failure of hair follicle eruption and

anaemia. Factors assumed to contribute to the anaemia include blood loss, impaired iron absorption and a reduction in the capacity of ascorbic acid to protect folate from irreversible oxidation. Patients with scurvy may also experience changes in personality and psychomotor performance: it is thought that these are due to impaired synthesis of catecholamines by dopamine β -hydroxylase, a vitamin C-dependent enzyme. Dietary deficiency can occur, particularly if food is overcooked, and requirements are increased in certain patients, for example following trauma or surgery or in the presence of certain drugs, for example sulfasalazine, aspirin. Smokers also seem to have an increased turnover of vitamin C.

There have been many attempts to demonstrate benefits from an excess intake of vitamin C (i.e. higher than that required to prevent scurvy) in a diverse range of conditions. None of these has yet been proven, although as one of the proponents, Linus Pauling, was a double Nobel laureate, they remain widely held beliefs. Potential risks of very high doses include diarrhoea and increased formation of urinary oxalate, although the latter may be an *in vitro* effect in stored urine, and thus not a risk factor for renal stones.

Other organic substances. Food contains many other organic substances besides those described thus far. There is, however, no convincing evidence that any of these are essential nutrients in man, although some (e.g. choline, pangamic acid, laetrile) have been proposed as such, and choline is now regarded as an essential nutrient in the USA. Carnitine, which is involved in the transport of acyl-CoA across the mitochondrial membrane in fatty acid oxidation, is synthesized in the liver in adults, but there may be a dietary requirement in premature infants.

Some organic (and inorganic) substances are deliberately added to foods. They include preservatives, flavourings, colourings, sweeteners and processing aids. In the UK, there are strict legislative controls over the use of such additives.

There has been much interest in recent years in foods that contain nutrients that have benefits beyond their purely nutritional properties (sometimes known as functional foods), and some of these are considered further in Chapter 11.

Other organic compounds found in foods (e.g. alcohol, caffeine, solanine) have potentially toxic pharmacological effects. Solanine is a plant glycoalkaloid that occurs in small amounts in all potatoes, but at much higher concentrations in green potatoes. It acts as a cholinesterase inhibitor, and ingestion of large amounts causes abdominal pain followed by nausea, vomiting and respiratory problems that can be fatal. Cholinesterase activity returns to normal within a few hours of consumption of small amounts of solanine, so there is not usually any cumulative ill effect from repeat ingestion of small doses. Other toxins are formed in food as a result of processing or cooking it: for example 3-monochloropropane 1,3-diol (from industrial processing of vegetable protein), acrylamide 2-propenamide (from frying or baking foods) and polycyclic aromatic hydrocarbons (in foods exposed to wood smoke) are all carcinogenic, at least in animals.

Trans fatty acids are formed during the hydrogenation of vegetable and fish oils, although they are also present naturally in meat and dairy products: an increased intake is associated with increased cardiovascular risk.

Contaminants in food have the potential to cause harm. Examples include bacteria, for example *Campylobacter jejuni*, which causes gastroenteritis; moulds, for example patulin, which can grow on apples, and aflatoxins from moulds found in foods, including peanuts and corn, which are carcinogens particularly associated with hepatic cell carcinoma.

Trace elements

Minerals are important nutrients, some being major body constituents while others (the trace elements) are required in much smaller amounts. The trace elements that are known to be essential (zinc, copper, selenium, molybdenum, manganese and chromium) are considered here, although there are a number of others (e.g. arsenic, cadmium, fluorine, lead, lithium, nickel, silicon, tin, vanadium) that have been suggested as having physiological roles, but which have yet to be proven to be essential.

Zinc. Zinc is an essential component of over 200 enzymes, having both catalytic and structural roles. Some enzymes appear to be more sensitive to zinc deficiency than others, and it is of relevance to the clinical biochemist that plasma alkaline phosphatase activity tends to decrease when zinc is deficient. Zinc is also incorporated into proteins that do not have enzyme activity; for example, it has a role in maintaining the aggregation of presecretory insulin granules and has direct and indirect roles in reducing free radical activity. Thus, zinc is important in most of the major metabolic processes in the body, including protein synthesis, cellular replication and collagen synthesis.

The best dietary sources of zinc are red meat, shellfish, nuts and cereals, although processing food tends to remove zinc (e.g. white flour is a poor source). Only about 30% of dietary zinc is absorbed and there is a significant enterohepatic circulation. Absorption occurs through both active and passive processes. Some zinc remains in enterocytes bound to a metallothionein, the proportion depending on the amount of metallothionein present. When the body content of zinc is high, metallothionein synthesis is stimulated, reducing the amount of zinc entering the circulation.

Although zinc is present in all body tissues, 60% is in skeletal muscle, 30% in bone and 4–6% in the skin. In the circulation, 80% is contained within red cells, while in the plasma, up to one-third is firmly bound to α_2 -macroglobulin and most of the remainder is loosely bound to albumin. The main route of excretion is in the faeces.

As might be expected from the diversity of its functions, deficiency of zinc tends to have a wide range of effects. Dermatitis and hair loss are frequent, together with growth retardation and poor wound healing. Anorexia, lethargy and impaired sensations of taste and smell may also occur. Immune function may be depressed, and since zinc has a role both in the hepatic synthesis of

retinol-binding protein and in the conversion of retinol to retinal in the retina, night vision may be impaired.

Zinc deficiency can arise from inadequate intake (e.g. during parenteral feeding), impaired absorption or increased losses (e.g. through a gastrointestinal fistula). Zinc deficiency is an important part of protein–energy malnutrition and may limit weight gain during refeeding unless adequate amounts are provided. Requirements for zinc are increased in premature infants, during pregnancy and lactation and following trauma. The autosomal recessive disease acrodermatitis enteropathica is due to severely impaired intestinal absorption of zinc. Affected children have a scaly dermatitis (often secondarily infected), diarrhoea and impaired growth, all of which respond to oral zinc supplements in relatively high doses.

Chronic ingestion of high zinc intakes in otherwise healthy people may lead to features of copper deficiency (e.g. microcytic anaemia, neutropenia). This is because the induction of metallothionein synthesis in enterocytes by zinc results in increased binding of copper to metallothionein, which prevents its transfer to the circulation. (This effect is utilized in the treatment of some patients with Wilson disease; see Chapter 14.) Acute toxicity has been recorded following ingestion of water with a high zinc content, or the use of such water in renal dialysis.

Copper. The main role of copper is as a component of copper metalloenzymes, of which there are many. In the synthesis of collagen and elastin, the cross-linking reactions require various copper-containing amine oxidases, and copper is also involved in the oxidation of Fe^{2+} to Fe^{3+} during haemoglobin formation and the binding of iron to transferrin. Examples of other copper-containing enzymes include superoxide dismutase, tyrosinase and cytochrome *c* oxidase.

Copper is widely distributed in the diet, occurring, for example in shellfish, liver and kidney, with lesser amounts in milk, meat and cereals and very little in processed foods. Not all dietary copper is absorbed, the proportion varying from 35% to 70% for reasons that are still not fully understood, although an increased intake of zinc reduces absorption (see above). From the gut, copper is carried to the liver bound to albumin and there is incorporated into caeruloplasmin. Caeruloplasmin is then secreted into the blood and accounts for 80–90% of the circulating copper. The main route for excretion of copper is in the bile, with very little in the urine unless renal damage is present or copper-binding substances (e.g. penicillamine) have been given. Normal copper homeostasis, therefore, depends on the balance between intestinal absorption and biliary excretion.

Overt copper deficiency is rare, since the liver contains substantial stores, but has been seen in malnourished children and in adults on long-term total parenteral nutrition (TPN). Features include neutropenia and skeletal fragility, with a microcytic hypochromic (iron-resistant) anaemia if the deficiency is prolonged. In adults, subclinical deficiency may be a risk factor for cardiovascular disease – in animals, copper deficiency can cause vascular and myocardial disease, together with hypercholesterolaemia. There is also a rare X-linked recessive disorder known as Menkes disease, in which there is impaired copper

absorption and renal copper wasting. Affected boys show progressive developmental and mental retardation, their hair becomes depigmented and sparse and their bones are abnormal. Parenteral copper can restore both circulating and tissue concentrations, but is otherwise ineffective and death typically occurs by the age of three.

Excess copper in the body is toxic, as demonstrated by Wilson disease, which is discussed in Chapters 13 and 14. However, copper overload purely from chronic dietary excess is rare, although it has been described in association with contaminated water supplies. Although a problem of toxicology rather than nutrition, acute ingestion of copper salts can cause nausea, vomiting and diarrhoea, with intravascular haemolysis in severe cases.

Selenium. Selenium is essential to the activity of the enzyme glutathione peroxidase, which is an important component of cellular antioxidant capacity. Selenium deficiency results in both reduced enzyme activity and reduced amounts of the enzyme protein, although when the intake of selenium is increased, there is an increase in glutathione peroxidase activity only up to a plateau level. Other selenoproteins are known, including iodothyronine deiodinase, which catalyses the conversion of thyroxine to tri-iodothyronine, and thioredoxin reductase, which reduces nucleotides in DNA synthesis.

Cereals, meat and fish contribute most of the dietary selenium, mainly in the form of the amino acids selenomethionine and selenocysteine. About half of dietary selenium is absorbed, the main route of excretion being in the urine. Selenium deficiency, when it occurs, is generally the result of a low intake and has been described in patients on long-term parenteral nutrition and children with certain inherited metabolic diseases treated with very restrictive therapeutic diets, although such occurrences are rare. In certain areas of China, where the soil selenium is low, there is endemic deficiency of selenium and a selenium responsive cardiomyopathy (Keshan disease) has been described. However, a low average selenium intake has also been reported in other countries, with no specific selenium responsive diseases being described. Nonetheless, more recent epidemiological evidence has suggested an increased incidence of cancer in low selenium areas, and some countries (e.g. Finland) now add selenium to the fertilizers used for arable crops.

Excess dietary selenium causes toxicity in livestock, but it is only in selenium-rich areas of China that convincing endemic dietary selenosis has been described in man. Features include 'garlic breath' (due to dimethyl selenide), deformity and loss of hair and nails, skin and gastrointestinal symptoms and, in severe cases, neurological problems. Similar features have been seen in people taking selenium supplements that accidentally contained vastly more selenium than they were supposed to. In general, selenium toxicity seems to begin at intakes only ten times the normal intake, although the threshold may be even lower in the presence of certain diseases, for example cystic fibrosis.

Molybdenum. Molybdenum combines with molybdopterin to form molybdenum cofactor, essential for the activity of the enzymes xanthine oxidase, aldehyde oxidase and sulphite oxidase.

The dietary requirement for molybdenum is very small and it is present in most human diets in meats, legumes and grains, so that dietary deficiency is exceptionally rare. Molybdenum deficiency has been reported in a patient with Crohn disease on TPN, who developed fatigue, somnolence and amino acid intolerance that responded to molybdenum supplements. There is a rare, autosomal recessive, inherited deficiency of molybdenum cofactor, which results in severe neurological abnormalities, mental retardation, lens dislocation and xanthinuria (see Chapter 9) in affected children, but no clinical improvement results from dietary supplementation.

High dietary intakes of molybdenum may be associated with altered purine metabolism (an increased incidence of gout has been noted in some populations) and with poorly understood changes in copper metabolism.

Manganese. Manganese is a component of certain enzymes (e.g. pyruvate carboxylase, mitochondrial superoxide dismutase, arginase) and is also an activator of many others (e.g. hydrolases, glycosyl transferases, kinases, decarboxylases), so deficiency could potentially affect the metabolism of carbohydrates, glycosaminoglycans and cholesterol.

Good dietary sources of manganese include leafy vegetables, unrefined cereals and tea, although only a small proportion (3–4%) is actually absorbed. The body content of manganese is low, with about a quarter of this relatively fixed in bone and the highest tissue concentrations in the pancreas and liver. The usual route of excretion is in the bile.

Deficiency of manganese in animals causes poor growth, defective collagen formation, depressed reproductive function, impaired glucose tolerance and neurological deficits, but the effects of dietary deficiency in man are less well established. A case of probable manganese deficiency occurred in a man on long-term TPN (who had a decrease in vitamin K-dependent clotting factors, hypocholesterolaemia, mild dermatitis and a slight change in hair colour). Dietary toxicity of manganese is unlikely since absorption is low and excretion in the bile and urine efficient. Toxicity has been recorded in manganese miners, who develop a condition clinically similar to Parkinson's disease, but this appears to be due to absorption from manganese ore dust in the lungs rather than from the intestine. High signals on T1-weighted magnetic resonance imaging (MRI) scans of the brain, consistent with manganese deposition, have been well documented in patients on long-term TPN in association with evidence of Parkinson disease and neuropsychiatric abnormalities.

Chromium. The main biological role of chromium in man appears to be to potentiate the action of insulin, as part of a low molecular weight chromium binding substance (chromodulin). Chromium may also be important in gene expression, lipoprotein metabolism and in maintaining nucleic acid structure.

Dietary sources of chromium include yeast, meat, whole grains, mushrooms and nuts. Absorption is very low and urinary excretion increases with intake. Human deficiency was first observed in long-term parenteral nutrition patients who developed insulin-resistant glucose intolerance

and neuropathy. In animals, deficiency also causes impaired growth and fertility and hypercholesterolaemia, but these have not been seen in humans. In general, there is no conclusive evidence that diabetes and hypercholesterolaemia are usually associated with chromium deficiency, nor that supplementation ameliorates these conditions.

As a normal dietary constituent, chromium toxicity seems unlikely to occur. However, the absorption of chromium picolinate, which is marketed as a dietary substitute, is higher than from other dietary sources and there have been reports of adverse effects, including renal failure. Industrial toxicity of chromium salts is well established and, recently, local and systemic effects of chromium and cobalt from metallic joint prostheses have been recognized.

Fibre

Dietary fibre is a term used to describe carbohydrate polymers that are not hydrolysed by the endogenous enzymes of the gastrointestinal tract. These include edible carbohydrate polymers naturally occurring in the food, carbohydrate polymers that have been obtained from food raw materials by artificial means and certain synthetic polymers. Lignin (a complex carbohydrate polymer containing aromatic residues) is also included in the definition. Dietary fibre is, therefore, a mixture of substances, including non-starch polysaccharides (NSPs), and starch and oligosaccharides that are resistant to digestion. Most dietary fibre consists of NSPs, and consideration of the NSP content of foods is probably the most useful approach to fibre, particularly since NSPs can be measured with more specificity than total dietary fibre.

Non-starch polysaccharides are a complex group of polymers that can be classified in a number of ways, for example according to the constituent monomers or whether the resulting polymer is soluble or not. Cellulose is a polymer of glucose but, unlike starches, is an unbranched β -1,4 glucan. It constitutes about a quarter of the NSP intake of an average UK diet and about half of the insoluble NSP. Other, non-cellulosic, polysaccharides tend to be made up of hexoses (e.g. glucose in the soluble glucans), pentoses (e.g. arabinose and xylose in the partly soluble arabinoxylans) or uronic acids (e.g. galacturonic acid in the soluble pectins).

The exact composition of dietary NSPs depends on their source. The richest sources are whole grain cereals, with wheat, maize and rice containing mainly insoluble forms, and oats, barley and rye a significant proportion of soluble NSPs. Fruit and vegetables have a higher water content than cereals and so the proportional NSP content tends to be lower, with pulses and nuts probably being the next best sources after cereals. Vegetables tend to contain roughly equal proportions of the soluble and insoluble fractions, but in fruits, this varies widely, with uronic acid-derived NSPs being the main soluble form.

A low dietary fibre intake has been associated with an increased incidence of constipation, diverticular disease, appendicitis, gallbladder disease and carcinoma of the large bowel, although the precise basis of these associations is difficult to work out.

Analysing the benefits of dietary fibre is complicated by the fact that its exact quantification in food

is technically difficult and that studies have tended to examine different types and combinations of fibre in different populations. A recent meta-analysis found an inverse association between the intake of dietary fibre, cereal fibre and whole grains and the risk of colorectal cancer, but no association with intake of fibre from fruit, vegetables or legumes. Possible mechanisms for this protective effect include an increase in stool bulk with dilution of faecal carcinogens and decrease in transit time with reduction in contact of carcinogens with the colorectal epithelium. Bacterial fermentation of fibre leads to the production of short chain fatty acids, which may have an additional protective effect by reducing intra-luminal pH. A similar inverse association between the intake of dietary fibre and breast cancer risk has been shown and inhibition of intestinal reabsorption of oestrogens with increased faecal excretion has been postulated as the basis of this.

Certain of the soluble NSPs have also been associated with short-term decreases in plasma glucose, insulin and cholesterol concentrations, but as yet there is no firm evidence either for a dietary lack of NSPs being implicated in the pathogenesis of type 2 diabetes or hyperlipidaemia, or for a role for dietary NSPs in the management of these conditions, other than the fact that a high dietary NSP intake tends to be a marker for a diet more in line with current 'healthy eating' concepts. Higher intake of dietary fibre is often associated with other lifestyle factors such as increased physical activity, lower alcohol and dietary fat consumption and lower prevalence of obesity, which may act as confounding factors in evaluating its role.

High dietary fibre intakes in infants and young children may displace energy-rich foods and restrict growth. In adults, high dietary intakes of NSP-rich foods are probably not harmful, although they may cause flatulence, and there is a possible increase in the risk of mechanical bowel problems such as colonic volvulus. Divalent cations can bind to certain constituents of NSPs, for example to uronic acid residues, or to associated phytates, but this does not seem to cause mineral deficiency on a mixed diet. However, caution should be exercised in recommending dietary fibre supplements, particularly those high in phytates, to populations with borderline mineral balance, for example unprocessed bran for the elderly.

ASSESSMENT OF NUTRITIONAL STATUS

General

It is important to be able to assess nutritional status, not only in large surveys attempting to define dietary reference values, but also to determine whether individual patients' nutritional needs are being met, to identify patients likely to have increased morbidity and mortality without nutritional support, and to monitor progress with time.

However, the wide range of individual nutrients, and the fact that nutritional status is reflected in all body compartments, not all of which are easily accessible to

measurement, means that precise assessments are difficult to make. The various techniques that are used include:

- clinical assessment
- dietary assessment
- measurement of anthropometric indices
- functional assessment
- laboratory-based techniques.

Clinical assessment

Evidence of nutritional disorders may become apparent during the normal clinical processes of history taking and physical examination. The history should ascertain normal food intake and dietary preferences, the patient's usual weight and any unexplained weight changes. Any difficulties in chewing or swallowing are important, together with gastrointestinal symptoms such as anorexia, nausea, vomiting or altered bowel habit. In the past medical history, any chronic disease may have effects on nutritional status, and surgery involving the gastrointestinal tract or associated organs may also be of importance. Chronic use of alcohol or drugs (therapeutic or otherwise) may have nutritional implications, and socio-economic factors are important in terms of availability of foodstuffs, adequacy of preparation and general support.

Physical examination may reveal clinical signs suggestive of generalized nutritional depletion (e.g. muscle wasting, oedema); evidence of deficiency of specific nutrients as described above may also be present.

Dietary assessment

Dietary assessment involves reviewing the intake of food, and its individual dietary components, and comparing the amount consumed with reference values to see whether deficiency or excess is likely. To be maximally accurate, this would involve weighing all food eaten and analysing its chemical composition, which is clearly impractical for clinical purposes. Hence, an approximation of food intake must be arrived at and reference then made to standard food tables to calculate nutrient content (see Ashwell, in Further reading, below).

Methods of dietary assessment can be divided into those that record current intake, recall past intake or estimate typical intake. Recording current intake can be achieved in a quantitative way in which all food is weighed or measured prior to being eaten. However, this is intrusive even for a short period and an alternative technique is for the patient to describe the amount of food eaten; photographic atlases are available to improve accuracy. Dietary recall methods involve asking the patient to recount everything consumed, usually in the previous 24h. Whilst these methods lack accuracy and tend to underestimate food intake they can reveal major areas of deficiency, or excess, requiring further evaluation. Food frequency questionnaires evaluate typical food consumption over a longer period. They can be particularly useful for population studies, for example assessing vitamin D consumption in the elderly.

TABLE 10.2 Body mass index (BMI)

BMI	Description	Grade of obesity or chronic energy deficiency	BMI is calculated from the formula
			$\frac{\text{weight in kg}}{(\text{height in metres})^2}$
<16	Chronic energy deficiency	III	
16–16.9	Chronic energy deficiency	II	
17–18.4	Chronic energy deficiency	I	
18.5–24.9	Desirable		
25.0–29.9	Overweight (or pre-obese)		
30.0–34.9	Obese	I	
35.0–39.9	Obese	II	
>40.0	Obese	III	

The 'desirable' and other ranges shown are for adult white individuals – the ranges are probably different for other racial groups, for example in Asians, the risks associated with being overweight begin to increase above 23 kg/m².

Anthropometric measurements

Clinical assessment can often indicate under- or overnutrition. However, a previously well nourished patient may be in a negative nutritional state for a long time before appearing clinically undernourished. Anthropometry comprises measurements of the human body and there are several anthropometric indices that are useful nutritional indicators as a baseline for initial assessment and as tools for ongoing monitoring.

Height and weight. In adults, body weight can be used as a measure of nutrition in various ways. An individual's weight can be compared against 'ideal' or desirable weight using formulae or tables, although it is now usual to calculate the body mass index (BMI) (see Table 10.2) in this context. The BMI can be used to grade the severity of obesity or of chronic energy deficiency but as it cannot distinguish between fat mass and lean mass it can be unrepresentative in patients with a large muscle mass, for example athletes, or in patients with an increased weight due to fluid retention. Variations in BMI between ethnic groups are also documented. In situations when measurement of height is not possible (e.g. immobility, spinal disorder), alternative indices such as the demispan – the distance between the index/middle finger web and the sternal notch – can be used to calculate BMI.

In ill patients, calculation of weight lost as percentage of usual weight can be useful, weight loss of more than 10% being an indicator of poor clinical outcome. However, determination of weight loss by history has been proven to be inaccurate.

In the UK, it is recommended that all patients are screened for nutritional risk at presentation to a doctor or other healthcare professional, and, in in-patient settings (both in hospitals and in nursing homes), at regular intervals thereafter. A suitable tool for this purpose is the Malnutrition Universal Screening Tool (MUST), which considers weight and unintentional weight loss, the time over which nutrient intake has been reduced and/or the

likelihood for impaired nutrient intake in the future (see Further reading, below, for more details).

In hospitalized patients, serial weight measurements are valuable in monitoring nutritional status, particularly where nutritional support is being provided. Acute changes are invariably due to changes in fluid balance (unless poor technique is responsible), but longer-term trends reflect tissue changes. As long as the patient is ambulant, weight measurements should be simple to obtain and record, although this may be surprisingly difficult to achieve on a busy hospital ward.

Nutrition is an important factor in growth and, in children, measurement of height and estimation of rate of growth is an important part of the assessment of nutritional status, in addition to measurement of weight.

Circumference measurements. In patients who are confined to bed, either because of debility or because movement is restricted by multiple fluid lines or splints, weight measurements are difficult to obtain unless the patient is in a special (and expensive) weighing bed. In these circumstances, some idea of tissue bulk may be obtained by a mid-arm circumference (MAC) measurement made with a tape measure. The measure should be placed round the non-dominant arm, at a point midway between the acromion (shoulder tip) and the olecranon process (point of the elbow). Age- and sex-related reference standards are available for the raw measurements, which include muscle, fat and bone or attempts can be made to refine the measurement to reflect muscle mass alone. The first stage is to correct for fat by including the triceps skinfold (TSF, see below) measurement and using the formula $(MAC - (\pi \times TSF))$ to derive the mid-arm muscle circumference. This can be further manipulated to give uncorrected muscle area, and finally an adjustment made for the contribution of the humerus, to give corrected muscle area. The latter can then be used to estimate total body muscle mass, although given the difficulties in making the two original measurements accurately, the margin of error by this stage is quite high. In the research environment, dual-energy X-ray absorptiometry (DEXA) is being used to make similar measurements of body composition.

While these calculated values may be useful in epidemiological surveys, in nutritional support the initial MAC measurement alone may help to indicate whether support is required, and serial measurements help to assess the success of treatment, although short-term changes are more likely to be due to imprecision in measurement than changes in tissue bulk. Measurements of calf circumference can be used in a similar way.

Another 'circumference' parameter that may be useful in assessing the health risks of moderate obesity is to measure waist and hip circumferences and calculate the waist:hip ratio. This reflects the distribution of fat rather than the degree of obesity; the risk of ischaemic heart disease and stroke rises sharply with waist:hip ratios of more than 1.0 for men and 0.8 for women. Even more simply, measurement of waist circumference alone can be used in this context, with a significant increase in cardiovascular risk being associated with waist measurements of over 102 cm in men and 88 cm in women (in Caucasian populations).

Skinfold thickness. Measurement of skinfold thickness, obtained using calipers, is useful in assessing and monitoring nutritional status in patients who cannot be weighed, and also has a place in epidemiological surveys. However, the technique is prone to large variations, both within and between observers. The imprecision arises in identification of the exact location for measurement; the way the skinfold is picked up; the way the calipers are placed on the fold; the compression of the fold by the calipers and the exact timing of the reading. Some improvement in performance may be achieved by taking the mean of three readings, usually on the left (or non-dominant) side. As with circumference measurements, the presence of oedema at the measurement site may be a further confounding factor.

A variety of sites have been used for skinfold measurement, but the most common are triceps, biceps, subscapular and suprailiac. Equations are available for calculation of total body fat from these measurements (usually for research purposes), but this assumes that subcutaneous fat reflects total body fat, which is not always the case: obese men tend to lay down more intra-abdominal fat than women, and visceral fat and subcutaneous fat have been shown to be biologically distinct. In clinical practice, body weight is more useful than skinfold thickness in the management of obesity, but in undernourished patients the latter may be useful. Age- and sex-related reference standards are published (e.g. for triceps skinfold thickness) and so measurement at presentation may identify severe malnourishment: a triceps skinfold of <5 mm almost always reflects low body fat stores. Serial measurements can help in monitoring nutritional support but, again, short-term changes are more likely to reflect imprecision in measurement than sudden changes in fat stores.

Functional assessment

The effects of nutritional status on certain aspects of bodily function can be used in the assessment of undernutrition, although with variable degrees of success. The best example is probably the effect of haematinic deficiency on red blood cell morphology where, for example a hypochromic, microcytic picture may be the first indication of iron deficiency and macrocytosis of vitamin B₁₂ or folate deficiency (see Chapter 26).

Functional tests of muscle mass have also been used, for example grip strength, isometric knee extension and response to electrical stimulation. However, while muscle strength correlates with muscle mass in normal subjects, there are many non-nutritional factors that can cause weakness in sick patients, and malnutrition alone has to be quite severe before strength diminishes. There is also the possibility that repeated measurements may have a training effect on the muscles involved.

Visceral protein is sometimes disproportionately decreased in protein deficiency and various measures of visceral function have been used in nutritional assessment. Two of these are discussed below.

Hepatic secretory proteins. The liver synthesizes most of the circulating plasma proteins (apart from immunoglobulins) and, in epidemiological studies, there

is a clear correlation between plasma concentrations of these proteins and other markers of malnutrition. For example, in adults who are otherwise well, a plasma albumin <35 g/L and a plasma transferrin <1.5 g/L usually indicate protein malnutrition. Since most of these proteins can be measured quite easily, there is a temptation to use them as nutritional markers in individual patients, but there are many pitfalls for the unwary.

Albumin is probably the most frequently measured plasma protein and low concentrations may reflect a deficiency in dietary protein intake. However, it must be remembered that decreased synthesis may be due to other factors, for example liver disease, and that the plasma concentration is also affected by fluid balance, loss of protein from the body, tissue catabolism and distribution of albumin across the various body compartments. It should also be remembered that, while the intravascular concentration of albumin is relatively high, more than half of the total mass of albumin is actually extravascular. It is noteworthy that in uncomplicated starvation, plasma albumin concentrations may remain normal for a relatively long period (the fractional catabolic rate decreases), whereas in sepsis, the concentration can fall significantly over a much shorter period.

Even in circumstances where the plasma albumin concentration does reflect nutritional status, the relatively long half-life (20 days) means that it does not respond to rapid changes and so measurements of other plasma proteins with shorter half-lives have been used. These include transferrin (9 days), prealbumin (1–2 days) and retinol-binding protein (10 h), but unfortunately, these all have similar drawbacks to albumin and are also affected by factors such as the acute phase response, oestrogen concentrations and factors relating to their own specific function (e.g. iron deficiency in the case of transferrin). Serial determinations may be of more use than single measurements, and measurement of a marker of the acute phase reaction (e.g. C-reactive protein) at the same time may help in interpretation. In clinical practice, measurement of these proteins rarely contributes to the management of individual patients, and failure to appreciate their limitations may lead to inappropriate decisions being made.

The immune response. Malnourished people are more susceptible to infections. Plasma immunoglobulin concentrations are generally maintained, but cell-mediated immunity may be impaired. The circulating absolute lymphocyte count is often low (<2.0 × 10⁹/L) in malnutrition, although this is a very non-specific finding.

Delayed cutaneous hypersensitivity testing against common allergens has been used to assess malnutrition, but many non-nutritional factors (e.g. infection, malignancy, radiation, surgery, drugs) affect the response and, in any case, the results may not be reproducible. In vitro tests of T cell function may be an alternative, but in clinical practice neither of these approaches is in common use.

Laboratory-based assessment of individual nutrients

Energy

Laboratory-based techniques are not in general use for the assessment of energy stores, whether of fat or

carbohydrate. However, there are techniques for the measurement of energy expenditure that may be of use in assessing how much energy to provide as part of nutritional support.

Direct calorimetry measures heat loss from the body, which can be used to derive the metabolic rate assuming that body temperature remains constant and no external work is performed. The technique involves the subject remaining in a special, insulated room equipped for heat exchange, so is not applicable in general clinical practice. Indirect calorimetry measures oxygen consumption and carbon dioxide production, and from these the respiratory quotient can be calculated, together with the energy expenditure (Fig. 10.7). This is the technique used by the ‘metabolic measurement carts’ that have been popular in some intensive care units.

The use of doubly-labelled water (containing the stable isotopes ²H and ¹⁸O) has produced new data on total energy expenditure in free-living individuals, but the technique is not suitable for the assessment of patients.

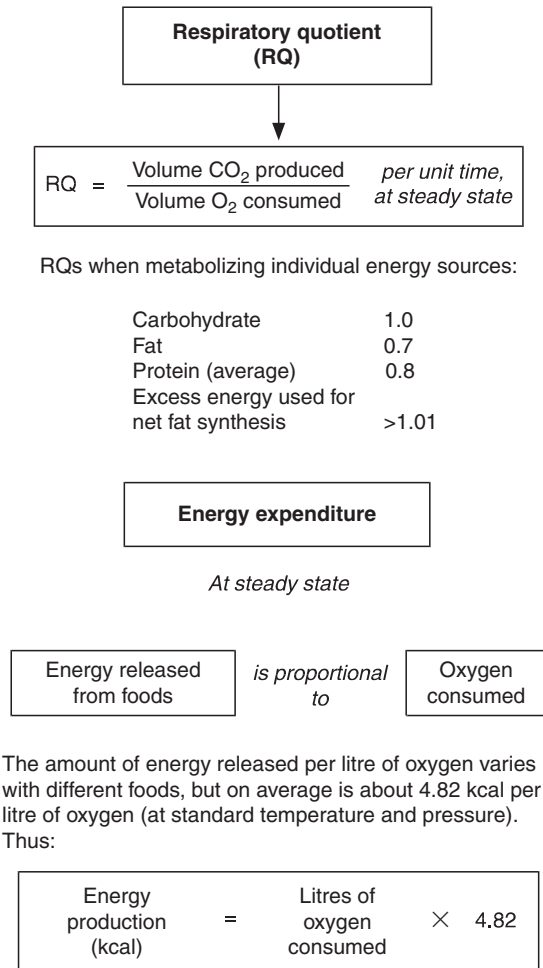


FIGURE 10.7 ■ Calculation of respiratory quotient and energy expenditure.

Heart rate can also be used as an index of energy metabolism, but although the measurement is obviously very much easier, the relationship between heart rate and energy expenditure is not linear and there are many confounding factors, so it can be useful in studies of groups of people but not in individuals.

Basal metabolic rate (BMR) can be estimated for healthy individuals from knowledge of age, sex, height and weight using the Harris–Benedict (Fig. 10.8), Schofield (Fig. 10.9) or similar equations and, for sick patients, suitable adjustment can then be made for disease state, pyrexia, mobility etc. However, the more severely ill the patient, the less likely it is that the calculated figure truly reflects total energy expenditure.

In disease-free individuals, the following formulae give an estimate of BMR in kcal/24h:

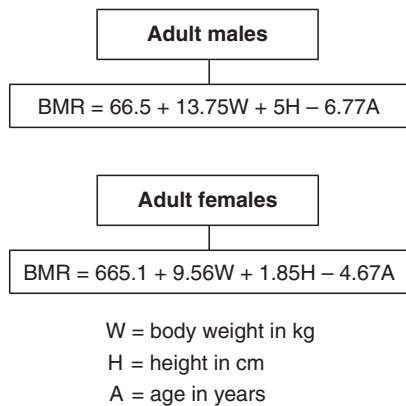


FIGURE 10.8 ■ Harris–Benedict equations for the calculation of basal metabolic rate (BMR).

As a simpler alternative to the Harris–Benedict equations, the following equations can be used to calculate BMR in kcal/24 h:

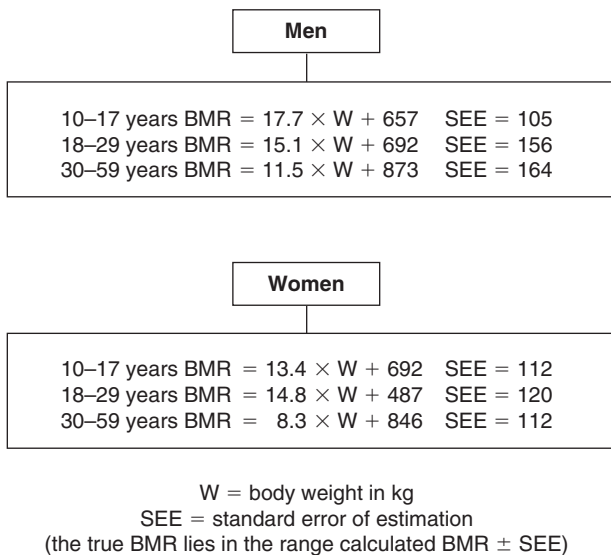


FIGURE 10.9 ■ Schofield equations for the calculation of basal metabolic rate (BMR) in kcal/24 h.

Protein

Laboratory-based tests for visceral protein have already been discussed under functional tests of nutritional status. However, there are a number of other tests that assess total body protein (predominantly muscle).

Total body nitrogen can be measured by neutron activation (an estimate of total body protein is simply obtained by multiplying this figure by 6.25) and this can demonstrate changes over a period of only a few weeks. However, the technique is expensive and involves a significant dose of radiation, so it is only used as a research tool. Almost all body potassium is intracellular; there is virtually none in adipocytes and potassium in the skeleton and red blood cells is not readily exchangeable, so exchangeable potassium measured using dilution of ⁴²K is a measure of fat-free mass. By monitoring both total body nitrogen and potassium, it can be shown that nitrogen losses in malnutrition are relatively higher from lean tissue (high potassium content) than from fibrous tissue (low potassium content).

Measurements of bioelectrical impedance have also been used to predict lean body mass. However, the technique actually measures total body water and the extension to lean body mass assumes a constant level of hydration in such tissue, which may not always be the case, particularly in sick individuals.

Outside the research environment, a direct measure of muscle mass may be obtained from the 24h excretion of creatinine, a metabolite of muscle creatine. This is sometimes corrected for height (the creatinine:height ratio) and standard values are available. However, urinary creatinine excretion does fluctuate (e.g. with exercise, pyrexia, meat intake) and the technique depends on complete urine collections and normal renal function, so that in practice, it is not widely used.

It may sometimes be useful to assess nitrogen balance. Nitrogen intake may require dietary analysis or, for patients on enteral or parenteral formulations, it will already be known. Nitrogen output for research purposes involves analysing stool and urine for total nitrogen content, but in the past, for clinical purposes, urinary urea nitrogen has been measured and standard adjustments made for urinary non-urea nitrogen and faecal nitrogen. This approach was valid for reasonably well patients (assuming the urine was not infected with a urea-splitting organism), but diverges further from the true nitrogen output of the more severely ill the patient, so its use cannot be recommended.

Vitamins

It is sometimes useful to assess body stores of fat-soluble vitamins, as outlined below, but only rarely necessary to measure the water-soluble vitamins (with the exception of vitamin B₁₂ and folate, see Chapter 26). This is because they are relatively non-toxic and, if deficiency is suspected, a trial of therapy is safe and simpler than trying to measure them. The interpretation of laboratory data is not without difficulty – some vitamins are affected by the acute phase response, for example vitamin D concentration may fall by up to 40% – so differentiation of true deficiency may be difficult. Measurement of concentrations of some vitamins in erythrocytes rather than plasma,

during the acute phase, has been shown to be more representative of body stores.

Vitamin A. The standardized definition of vitamin A deficiency is in terms of liver stores of retinol, but this is not a practical indicator. By convention, a plasma retinol concentration $<0.7 \mu\text{mol/L}$, is used to define deficiency. However, plasma retinol concentration tends to be maintained over a wide range of liver reserves, making it an insensitive detector of deficiency and as retinol is carried bound to retinol-binding protein (RBP) in the plasma, conditions reducing RBP concentration (e.g. liver disease) may lower plasma retinol concentration without reflecting decreased stores. The diagnosis of vitamin A deficiency is often made on clinical grounds. The response to a test dose of vitamin A may help to distinguish true deficiency from functional deficiency due to a lack of RBP. In the former, the plasma vitamin A concentration rises markedly within a few hours; in the latter it does not. Very high retinol intakes may increase the plasma concentration.

Plasma carotenes can also be measured and, in general, reflect dietary intake. Dietary assessment is usually all that is necessary for this, but measurement is occasionally helpful to establish whether an odd skin colouration is due to hypercarotenaemia or not.

Vitamin D. Patients with a nutritional deficiency of vitamin D may present with clinical symptoms and signs; suspicion may also be raised by the finding of an increase in plasma alkaline phosphatase activity. Plasma calcium may be low but is frequently within the reference range as a result of secondary hyperparathyroidism, which may result in a low plasma phosphate concentration.

The main storage form of vitamin D in the body is 25-hydroxycholecalciferol and this is also the main form in plasma, where it is bound to a specific binding protein. Measurement of plasma 25-hydroxycholecalciferol provides a useful indicator of vitamin D stores: the reference range for total 25-hydroxycholecalciferol is of the order of 50–140 nmol/L, although the result should be interpreted with regard to the time of year, as there can be a two-fold seasonal variation in temperate regions. Concentrations $<50 \text{ nmol/L}$ indicate vitamin D insufficiency that may benefit from supplementation, and values $<25 \text{ nmol/L}$ suggest deficiency.

It is also possible to measure the active metabolite of vitamin D (1,25-dihydroxycholecalciferol, calcitriol), although this is more difficult as the plasma concentrations are approximately 1000-fold lower than those of 25-hydroxycholecalciferol. Its synthesis is normally closely regulated and concentrations do not reflect vitamin D status. Such measurements tend only to be necessary in the investigation of rare, inherited disorders of vitamin D metabolism or action (see Chapter 30).

Vitamin D toxicity can occur if supplements are given, as the difference between appropriate and toxic intakes is fairly small. This is generally detected in the laboratory as hypercalcaemia that gets better when the supplements are reduced or stopped, but it is occasionally necessary to confirm the diagnosis with plasma vitamin D measurement.

Vitamin E. Vitamin E status is assessed by measurement of plasma tocopherol concentration, most commonly by high performance liquid chromatography. Increases in plasma lipoprotein concentrations tend to result in increases in plasma tocopherol, so it may be appropriate, in some circumstances (e.g. in patients with cholestasis or if the plasma tocopherol is found to be low in order to confirm deficiency) to express the result as a plasma tocopherol:cholesterol ratio.

Vitamin K. Although it is possible to measure vitamin K in plasma, the usual laboratory approach is to measure the plasma concentrations of the vitamin K-dependent clotting factors. In vitamin K deficiency, these are present in normal quantities, but in a non-functional state (proteins formed in vitamin K absence – PIVKA), and so a functional assay, conventionally the prothrombin time or international normalized ratio (INR), is used (although the activated partial thromboplastin time is also abnormal in vitamin K deficiency).

There is no generally available test of vitamin K excess, largely because its lack of toxicity makes it unnecessary.

Thiamin. The single most important factor in detecting thiamin deficiency is a high index of clinical suspicion, and the safest confirmatory test is a trial of thiamin supplements, since delay can be harmful in a deficient patient. However, assuming appropriate samples can be obtained without delaying treatment, there are laboratory tests that can be used to assess thiamin status.

Methods are now available for measurement of thiamine and its phosphate esters in whole blood, but because of the low concentration and consequent analytical problems, some laboratories still use functional tests of thiamin deficiency, the most common of which is probably measurement of red cell transketolase with and without the addition of thiamin pyrophosphate (TPP). In normal subjects, TPP enhancement of transketolase is up to 15%, but may be increased to as much as 25% in severe deficiency. Urinary thiamin measurement has been used, with low outputs being found in clinical deficiency states. Discrimination is said to be improved if the percentage excretion of a test dose is measured; a higher proportion is retained in deficiency.

Riboflavin. The laboratory approach to the assessment of riboflavin (vitamin B₂) status is similar to that taken with thiamin. Urinary excretion can be measured and an output of 0.5–0.8 mg/24 h probably indicates an adequate intake, but does not necessarily reflect total body content. Factors such as age, physical activity and illness can also influence urinary excretion. Discrimination may be improved by measuring the percentage excretion of a loading dose of riboflavin, an increased proportion being retained in deficiency states.

In the blood, about 90% of riboflavin is present in the red cells. This can be measured directly, but measurement of red cell glutathione reductase activity is probably more useful. The enzyme is flavin adenine dinucleotide (FAD) dependent, so that activity falls in riboflavin deficiency, but is restored on addition of FAD. An increase in activity of 30% or more confirms riboflavin deficiency. The test assesses long-term tissue riboflavin status and is

very sensitive. It cannot be used if there is coexistent red cell glucose 6-phosphatase deficiency.

Nicotinamide. Laboratory tests of nicotinamide status tend not to be entirely satisfactory, possibly because no specific functional test has yet been developed. Approaches used to date include measurement of whole blood NAD concentration and of urinary excretion of nicotinamide metabolites, in particular 1-methyl nicotinamide.

Vitamin B₆. A variety of biochemical techniques have been described for the assessment of vitamin B₆ status. Urinary excretion of 4-pyridoxic acid reflects immediate dietary intake and pyridoxal 5-phosphate in blood can also be measured, but indirect approaches are less technically demanding and have been more widely adopted.

In vitamin B₆-deficient subjects, the urinary excretion of xanthurenic acid shows a marked increase following an L-tryptophan load, and this is the basis of the tryptophan load test. This is relatively easy to perform and has been widely used, but interpretation of the results requires caution as there are a number of other factors that can affect tryptophan metabolism. The most useful test currently available is probably the measurement of red cell aminotransferase (e.g. aspartate aminotransferase) activity, with and without the addition of pyridoxal 5-phosphate, in a technique analogous to that used for the functional assessment of thiamin (red cell transketolase) and riboflavin (red cell glutathione reductase). Serum aminotransferase activities are also affected by vitamin B₆ deficiency, but the very wide fluctuations seen in disease states make them less useful than erythrocyte activities in this context.

Pantothenic acid. Tests of pantothenic acid status are rarely necessary and no biochemical test has become generally accepted. Pantothenic acid itself can be measured in urine and serum, the urinary excretion being more immediately affected by dietary intake. Assessment of the red cell content of coenzyme A may be a functional test of pantothenic acid status.

Biotin. Techniques have been described for the measurement of biotin in both urine and blood. However, these are not straightforward and, in practice, biotin status is rarely assessed biochemically.

Vitamin C. Vitamin C can be measured in plasma, white blood cells and urine. Unfortunately, none of these is entirely satisfactory in assessing vitamin C status. Plasma concentration tends to reflect recent intake and so will fall with decreased intake before body stores are depleted. Leukocyte (or buffy coat) vitamin C content probably reflects body stores better than the plasma concentration, but the measurement is technically difficult, requires large quantities of blood and, if the buffy coat content is measured, any abnormality in the relative proportions of white cells and platelets will introduce further error.

Urinary ascorbate can be measured relatively easily in a fresh sample, and output decreases quite markedly in deficiency. However, the technique is not very

specific and various ascorbic acid saturation tests have been devised to improve on this. These all assume that if body stores are depleted, most of a standard daily dose of ascorbic acid will be retained until stores are repleted, after which relatively large amounts of ascorbate appear in the urine. These tests are not perfect, but are simple to carry out and appear to be satisfactory if a clinical suspicion of vitamin C deficiency needs to be confirmed. An added advantage is that any deficiency is actually treated by the test.

Trace elements

There is a general problem in assessing body stores of trace elements, in that plasma concentrations are not necessarily related to tissue content, as is apparent in the following discussion.

Zinc. Assessment of body zinc status is not easy. Measurement of serum zinc concentration is probably the simplest approach, but the result is influenced by many things other than total body zinc. There is a diurnal variation with a peak at about 10.00 h, fluctuation with recent dietary intake and a fall with hypoalbuminaemia. It also tends to fall in disease: for example, in acute inflammation, interleukin-1 stimulates zinc uptake by the liver, although in tissue destruction, the zinc may be maintained as zinc is released from the tissues, even though net balance is actually negative as losses in urine increase.

Urinary zinc excretion is also difficult to interpret. Low values may reflect deficiency, but, even if environmental contamination can be avoided, normal or raised excretion does not exclude deficiency; urinary losses may even be the cause of the deficiency.

Since serum and urine zinc measurements are generally unsatisfactory, a variety of other approaches have been tried. Hair zinc is liable to environmental contamination (e.g. from certain shampoos), turns over relatively slowly and depends on the rate of hair growth. Erythrocyte zinc content responds only slowly to changes in zinc status, which can be rapid. (However, the high zinc content of red cells, in the enzyme carbonate dehydratase, is important, since it means that *in vitro* haemolysis increases the measured serum zinc concentration.) Changes in the plasma and tissue activities of zinc-dependent enzymes like carbonate dehydratase and alkaline phosphatase have given varying results. Leukocyte zinc content correlates well with muscle content, but its measurement is not easy and requires large quantities of blood.

Serum zinc remains a common, albeit unsatisfactory, measure of zinc status, although some authorities state that a serum zinc concentration <7 µmol/L is a clear indication of deficiency. When true deficiency is suspected, a therapeutic trial of zinc supplementation with careful clinical monitoring is probably the best approach.

Copper. The diagnosis of copper overload is discussed in Chapters 13 and 14. In the detection of copper deficiency, plasma copper concentration may be low, but deficiency may be masked because caeruloplasmin synthesis is increased in the acute phase response. Caeruloplasmin

is also increased in pregnancy and by exogenous oestrogens. Measurement of caeruloplasmin before and after moderate copper supplementation may be a useful technique; an increase confirming deficiency. Interpretation of copper concentrations in other body fluids and tissues is difficult when looking for deficiency, although changes in the activities of copper-containing enzymes like superoxide dismutase may prove to be useful functional tests of copper deficiency in the future.

Selenium. Plasma selenium concentration is an indicator of recent selenium intake, rather than body stores. Measurement of whole blood selenium concentration has been used but may give a false picture if there has been recent blood transfusion. Selenium in plasma is bound to lipoproteins, which may decrease in concentration in malnutrition, causing a secondary decrease in plasma selenium not necessarily related to body content. The activity of the selenium-containing enzyme glutathione peroxidase tends to decrease in selenium deficiency and may be a better functional indicator, but it is not clear whether the measurement is best made in erythrocytes, platelets or even plasma. Excess selenium is excreted in the urine, so urinary measurements are useful in suspected toxicity, but are not helpful in deficiency.

Molybdenum. Plasma and urine concentrations of molybdenum are normally very low and measurements in these fluids are not useful in detecting deficiency. Increased urinary excretion of xanthine and sulphite, which decrease with molybdate supplementation, may be a useful approach in confirming suspected deficiency.

Manganese. Measurement of plasma and urine manganese is useful when investigating suspected toxicity, but since most manganese is intracellular, whole blood measurements are used in the detection of deficiency.

Chromium. Plasma and urine chromium can be measured by inductively coupled plasma mass spectrometry, but the limit of detection is such that they are more useful in detecting toxicity.

CONCLUSION

Clinical nutrition is a huge topic, with many different professions and specialties involved. Biochemical tests have their limitations in measuring nutritional status, and it is important for the clinical biochemist to be aware of these, in order that the tests are not used inappropriately.

ACKNOWLEDGEMENT

The author would like to thank Stephen K. Bangert who wrote the chapter in the previous edition of this book.

Further reading

Ashwell M, editor. McCance and Widdowson. A scientific partnership of 60 years. London: The British Nutrition Foundation; 1993.

Fascinating accounts of the early work on nutrition by McCance and Widdowson.

Department of Health. (Report on Health and Social Subjects 41) Dietary reference values for food energy and nutrients for the United Kingdom. Report of the panel on dietary reference values of the committee on medical aspects of food policy. London: HMSO; 1991.

Still in print and a useful starting point for information about individual nutrients.

Geissler C, Powers H, editors. Human nutrition. 12th ed. Edinburgh: Churchill Livingstone; 2011.

The twelfth edition of the classic text Human Nutrition and Dietetics, previously edited by Garrow, James and Ralph.

Malnutrition Universal Screening Tool (MUST).

This is a five-step screening tool to identify adults who are malnourished, at risk of malnutrition, or obese. It can be downloaded from the website of the British Association for Parenteral and Enteral Nutrition: www.bapen.org.uk. Accessed November 2012.

Nutritional disorders and their management

Ruth M. Ayling

CHAPTER OUTLINE

INTRODUCTION 200

MALNUTRITION 200

Protein–energy malnutrition in children 201
Chronic energy deficiency in Western adults 201

OBESITY 201

Aetiology of obesity 202
Secondary causes of obesity 203
Appetite 203
Management of obesity 205

EATING DISORDERS 207

Anorexia nervosa 207
Bulimia nervosa 208

DIET IN THE AETIOLOGY OF DISEASE 208

Dental caries 208
Cancer 208

THERAPEUTIC DIETS, DIETARY SUPPLEMENTS AND NUTRACEUTICALS 208

PROVISION OF NUTRITION SUPPORT 209

Indications for nutrition support 209
Enteral feeding 210
Parenteral nutrition 211
Short bowel syndrome 213

CONCLUSION 213

INTRODUCTION

The term ‘nutritional disorders’ covers a wide range of conditions. Some of these disorders are primarily nutritional and some are not, albeit that nutrition is still an important factor. This makes their classification somewhat arbitrary, but the following broad categories can be recognized:

- specific deficiency caused by inadequate supply of individual nutrients and occasionally, oversupply of an individual nutrient leading to toxicity, which has been discussed in Chapter 10
- generalized undernutrition, which is an important cause of morbidity and mortality across the world
- generalized overnutrition, in the form of obesity, which has become an important public health problem
- eating disorders are not primarily nutritional disorders, but have important nutritional effects and significant metabolic consequences
- specific diseases in which dietary factors may be important in their aetiology. In some cases there may be scope for dietary modification to alter the course of the disease (e.g. hyperlipidaemia). In others, this is not so (e.g. where carcinogenesis has already been stimulated by a dietary component)
- conditions in which diet has no role in aetiology, but for which specific dietary intervention

is an essential part of management (e.g. phenylketonuria)

- primarily non-nutritional diseases in which general nutritional support may improve outcome, especially if patients are, or are likely to become, malnourished during the course of their illness.

MALNUTRITION

‘Malnutrition’ is often taken to mean inadequate nutrition, but this definition also includes dietary excess, so the term ‘undernutrition’ is to be preferred. Undernutrition is a major public health problem in the developing world. Diets of affected populations tend to be deficient in both macronutrients and micronutrients. A high prevalence of bacterial and parasitic disease tends to contribute to undernutrition; in addition, an effect of undernutrition is to increase susceptibility to infection. In children, severe undernutrition is predominantly a state of protein energy malnutrition (PEM). In adults, in whom protein requirements are relatively lower (since growth has ceased), the term chronic energy deficiency (CED) is used. In the UK, severe undernutrition is most commonly seen in adults, usually in association with intercurrent illness and is discussed later in the chapter under nutrition support.

Protein–energy malnutrition in children

Worldwide, PEM remains the most common disorder of infancy and childhood: as many as 200 million children may be significantly affected at any one time and it accounts for over half the deaths in the under-five age group in developing countries. One way of assessing and categorizing undernutrition is by using a weight for height z score, which compares a child's height to that of a healthy reference population of children of the same height or length. The z score is expressed in terms of standard deviations (SD) from the mean of the reference population. When height or length cannot be measured easily, upper arm circumference, which changes little in healthy children between the ages of six months and five years, can be used as a measure of lean body mass to identify undernutrition. Using these methods, acute undernutrition can be divided into moderate and severe. Moderate undernutrition is defined as a weight for height or length z score between -2 and $-3SD$. Severe undernutrition is defined by a z score of $<-3SD$, an arm circumference <110 mm or the presence of nutritional oedema. Moderate or severe undernutrition without nutritional oedema is termed marasmus. Undernutrition with oedema is termed kwashiorkor.

Marasmus tends to occur in younger children who have the greatest energy requirements and susceptibility to infection. Affected children appear emaciated with loss of subcutaneous fat, muscle wasting and triangular facies due to loss of buccal fat pads. In addition to oedema, children with kwashiorkor have enlarged and fatty livers. They may also have dermatoses and depigmentation of hair, the latter being attributed to reduced availability of tyrosine, a precursor of melanin.

The classic explanation for the two different syndromes suggests that marasmus occurs on a diet poor in protein and energy and kwashiorkor on a diet poor in protein but with relatively more carbohydrate. In marasmus, insulin secretion would, therefore, be expected to be low, prompting breakdown of muscle protein and mobilization of amino acids for hepatic synthesis of protein, particularly albumin. In kwashiorkor, relatively higher carbohydrate intake would lead to maintenance of insulin secretion with relative sparing of muscle protein, therefore limiting amino acids available for synthesis of liver proteins. Reduced synthesis of albumin and apolipoproteins would lead to hypoalbuminaemia and oedema, and accumulation of lipids in the liver, respectively. However, recent studies have found marasmus and kwashiorkor to exist in populations where children are eating similar diets. It has been suggested that the clinical syndromes may in fact reflect differences in the metabolic response to starvation, for example varying rates of fatty acid oxidation, rather than the content of the diet.

The treatment of PEM is the provision of appropriate nutrition and treatment of associated infection. Treatment of moderate undernutrition involves the addition to the diet of suitable nutrient rich supplements containing significant energy and the recommended daily allowance of micronutrients; examples include cereal and legume blended flours and lipid rich fortified spreads.

Severe PEM is associated with atrophy of the small bowel mucosa, reducing its capacity for digestion and absorption, hence liquid, milk-based products, supplemented with potassium and phosphate, are favoured for treatment, initially administered as small boluses. Intravenous fluids should be avoided, except in children with profuse diarrhoea. Children with kwashiorkor have increased total body water with sodium retention. Those with marasmus have chronic hypovolaemia with secondary hyperaldosteronism, although tissue breakdown and urinary potassium loss tend to result in hypokalaemia. However, children with marasmus are particularly likely to have a degree of cardiac atrophy and reduced stroke volume, predisposing to cardiac failure if fluid overloaded. In addition, the reduction in subcutaneous fat renders these children susceptible to hypothermia and hypoglycaemia during the recovery process.

Chronic energy deficiency in Western adults

Studies of both medical and surgical inpatients in developed countries indicate that a significant proportion of patients have CED, although variation in the criteria used to define this means that the exact prevalence is uncertain. Caution is required in applying the marasmus/kwashiorkor distinction to such patients, since the presence of hypoalbuminaemia or oedema is more likely to be due to the underlying illness than malnutrition. Nonetheless, severe CED in such patients has important effects. Wasting of respiratory muscles increases the risk of chest infection and may delay weaning from a ventilator. Myocardial function may be impaired and skeletal muscle wasting delays mobilization, with a consequent increased risk of thromboembolism and bed sores. Chronic energy deficiency also results in impaired resistance to infection, and gut permeability may be increased, allowing entry of bacteria and toxins through the gut wall. Apathy and depression impair active efforts at recovery and may also impair appetite, worsening the situation further.

The recognition of undernutrition in adults and its treatment are considered later in this chapter in the section on nutritional support.

OBESITY

Obesity is an excess of body fat. For epidemiological studies and population surveys, as well as clinical assessment of individual patients, body mass index (BMI) is frequently used to define overweight (BMI ≥ 25 kg/m²) and obesity (BMI ≥ 30 kg/m²). Whilst it is a useful index, BMI is not without problems, but alternative methods are expensive and not readily available. In children, there are marked changes in BMI with age. It is recommended that obesity in children is defined with reference to BMI centiles, and charts are available for this purpose, but there is no universally accepted BMI classification system. In the UK, it is suggested that the 91st and 95th centiles should be used in children as indicators of overweight and obesity, respectively. Estimates of the prevalence of obesity

are likely to be subject to variation because of differences in the definition, cut-offs and reference standards used.

It is estimated that at least one billion people worldwide are overweight or obese and at least 300 million are obese. In the UK, from 1993 to 2010, the proportion of men and women who were overweight remained stable at 40% for men and 30% for women. However, there was a gradual increase in the proportion classified as obese from 13.2% of men (16.4% of women) in 1993 to 26.2% of men (26.1% of women) in 2010. Obesity is associated with significant health risks and the economic cost has been estimated as 3–8% of total healthcare expenditure. There is a relationship between mortality risk and increased BMI which, although strongest in those below 50 years of age, persists into the ninth decade of life. In addition, various complications are associated with obesity (see [Table 11.1](#)).

Aetiology of obesity

The aetiology of obesity is complex and influences on body fat content include genetic, perinatal, behavioural

and environmental factors. Epidemiological data indicate that the consumption of a diet that is high in fat and the frequent consumption of 'fast food' increase the risk of obesity. This is well demonstrated by the recent rise in obesity in developing countries, whose populations have replaced their indigenous diet with a more energy-dense Westernized alternative. There is also a close relationship between low levels of physical activity and weight gain and the duration of television viewing is a predictor of obesity risk in both adults and children.

Family studies have established that hereditary influences are important in determining body weight, although, to-date, genetic studies have concentrated on severe, early-onset forms of obesity, which are rare. Mutations resulting in leptin (see below) and leptin-receptor deficiencies have been described. Affected children are of normal birth weight, but exhibit extreme hyperphagia and gain weight rapidly. Children with congenital absence of pro-opiomelanocortin (POMC) gene products also become obese, despite glucocorticoid deficiency, an endocrine state normally associated with weight loss. These patients characteristically have pale skin and red hair, resulting from an absence of the action of POMC-derived peptides on melanocytes. A number of heterozygous point mutations in POMC have been found, which increase the risk of obesity, although are not invariably associated with it. Heterozygous mutations in the melanocortin receptor MC4R have been associated with dominantly inherited obesity and have been found in 1–2.5% of those with a BMI of $>30\text{ kg/m}^2$.

There are a number of conditions which have obesity as a feature of their clinical phenotype. The majority of these are associated with short stature; this contrasts with simple obesity where children tend to be tall. Examples of pleiotropic obesity syndromes are shown in [Table 11.2](#).

TABLE 11.1 Complications associated with obesity

System	Complication
Cardiovascular	Coronary heart disease, cardiac failure, hypertension
Respiratory	Sleep apnoea, dyspnoea
Gastrointestinal	Fatty liver, reflux oesophagitis, gallstones
Musculoskeletal	Arthritis, back pain
Endocrine and metabolic	Type 2 diabetes mellitus, insulin resistance, polycystic ovary syndrome, amenorrhoea, dyslipidaemia
Skin	Acanthosis nigricans, intertrigo

TABLE 11.2 Pleiotropic obesity syndromes

Syndrome	Genetic basis	Clinical features
Alstrom	Autosomal recessive <i>ALMS</i> gene located at 2p13	Obesity and type 2 diabetes mellitus Blindness Dilated cardiomyopathy Sensorineural hearing loss Other: hepatic, pulmonary, renal and urological dysfunction
Bardet–Biedl	Autosomal recessive 14 genes code for BB proteins involved in ciliary action	Obesity Hypogonadism Intellectual impairment Polydactyly Retinitis pigmentosa
Carpenter	Autosomal recessive <i>RAB23</i> gene located at 6p11	Obesity Cardiac defects Craniosynostosis Polysyndactyly
Prader–Willi	15q 11–13 Imprinted sequence	Obesity and hyperphagia Hypogonadotrophic hypogonadism Hypotonia Intellectual impairment Short stature Small hands and feet

Secondary causes of obesity

Investigation of patients to exclude secondary causes of obesity is usually unrewarding, but remains important. Clinical features suggesting endocrine pathology should be pursued, but in-depth dynamic function testing to exclude rare disorders is not usually necessary. Hypothyroidism should be excluded by measuring serum thyroid stimulating hormone and free thyroxine concentrations. The investigation of possible Cushing syndrome may be more difficult, as false positive results have been described in pseudo-Cushing syndrome due to obesity. Growth hormone deficiency may also lead to weight gain, although morbid obesity is not usually present. Associations are also recognized between obesity and polycystic ovary syndrome, hypogonadism and insulinoma.

Some drugs are associated with weight gain. The effect is assumed to be due either to central effects on appetite (e.g. certain anticonvulsants and antipsychotics) or to peripheral metabolic effects (e.g. oral hypoglycaemic agents, protease inhibitors).

Appetite

Energy homeostasis is the result of a complex, integrated process involving neural, humoral and psychological factors that are integrated by the brain to ensure that nutrient supply remains at appropriate levels for different environmental conditions. In the past, there has not tended to be an evolutionary advantage in weight loss, hence the mechanisms underlying energy homeostasis are tailored to ensure a powerful drive to eat, especially after weight loss.

Central appetite control

The hypothalamus. The body's appetite control centres are located in the hypothalamus. The ventromedial hypothalamus is thought to act as the satiety centre and the lateral hypothalamus as the feeding centre. Histochemical and molecular imaging techniques have shown that the hypothalamic nuclei involved in feeding regulation form interconnected circuits, utilizing neuropeptides as well as the classic amine neurotransmitters

The hypothalamus receives both neural and humoral input. At the base of the third ventricle, the arcuate nucleus (Arc; also known as the infundibular nucleus in humans) is able to receive signals from the circulation. The blood-brain barrier (BBB) surrounding the Arc is not complete and this allows hormones such as leptin, secreted by adipocytes, and insulin, from the pancreas, to gain access to the afferent signalling pathway that regulates appetite.

From the Arc, monosynaptic projections are made to many other brain regions, with the projections to the paraventricular nucleus (PVN) being of importance in the regulation of food intake. Integration of peripheral and central signals relating to energy homeostasis takes place in the Arc. Two neuron populations play a role. The orexigenic (appetite-stimulating) peptides neuropeptide Y (NPY) and Agouti-related peptide (AgRP) are co-localized, while in another population of neurons, the anorectic (appetite-inhibiting) peptide cocaine- and amphetamine-regulated

transcript (CART) and pro-opiomelanocortin (POMC, the precursor of α -melanocyte stimulating hormone (α MSH)) are also co-localized. The orexigenic NPY/AgRP neurons inhibit the anorexigenic POMC neurons through γ -aminobutyric acid (GABA)-ergic interneuronal connections. Fasting increases NPY and AgRP, while CART and POMC expression is reduced.

The brain stem. The vagus nerve and sympathetic fibres transmit satiety signals from the gut and liver to the nucleus of the tractus solitarius (NTS) in the brain stem. In areas such as the area postrema, which is in the brain stem adjacent to the NTS at the base of the fourth ventricle, the BBB is incomplete, also allowing circulating hormones access to the brain. The brain stem and hypothalamus are linked by projections from the NTS to the PVN and lateral hypothalamus, and from the raphe nuclei to the arcuate nucleus. Neurons such as those expressing glucagon-like peptide-1 (GLP-1) receive afferents from the vagal and glossopharyngeal nerves, integrating and relaying sensory information to hypothalamic and brain stem centres. Whilst the brain stem plays a role in regulating the size of individual meals, it is thought that the hypothalamus is necessary for long-term energy balance and appetite regulation.

Hypothalamic messengers

Neuropeptide Y. Neuropeptide Y is a 36-amino acid peptide member of the family of peptides comprising NPY, pancreatic polypeptide (PP) and peptide YY (PYY). Neuropeptide Y is one of the most potent stimulators of feeding. At least five distinct G-protein coupled receptors (Y1, Y2, Y4, Y5 and Y6) mediate the actions of NPY, PYY and PP. In rodents, repeated administration of NPY leads to hyperphagia and obesity associated with decreased thermogenesis in brown adipose tissue, hyperinsulinaemia, hypercortisosteronaemia, reduced plasma testosterone concentrations, and insulin resistance in skeletal muscle.

Agouti-related protein. Agouti-related protein increases food intake by acting as an antagonist at central melanocortin-3 and melanocortin-4 receptors (see below). In contrast to the fairly short-lived effects of NPY, central administration of a single dose of AgRP to rodents leads to an increase in food intake for up to one week. Repeated administration leads to hyperphagia and obesity.

Melanocortins. Melanocortins are peptides that are cleaved from the POMC precursor molecule by tissue-specific post-translational cleavage (e.g. in the anterior pituitary, POMC gives rise to adrenocorticotrophic hormone, ACTH). They bind to a family of melanocortin receptors (MC1-R to MC5-R). In the arcuate nucleus, α MSH is released from POMC-expressing neurons that project to the PVN, where it acts through MC3-R and MC4-R to inhibit food intake. The endogenous antagonist AgRP (see above) is released from the terminals of arcuate NPY/AgRP neurons at the PVN: AgRP stimulates food intake by blocking the anorectic effect of α MSH. The MC4 receptor, in particular, seems to be critical to regulation of body weight; thus far, the most common cause of monogenic obesity in humans is mutation in MC4-R.

Cocaine- and amphetamine-regulated transcript. Cocaine- and amphetamine-regulated transcript is co-expressed with POMC in arcuate neurons, and these neurons are directly stimulated by leptin (see later). Central administration of CART to rats inhibits feeding and completely blocks the feeding response stimulated by NPY. Food-deprived animals show decreased expression of CART mRNA in the arcuate nucleus. CART may thus be another endogenous inhibitor of food intake.

5-Hydroxytryptamine. 5-Hydroxytryptamine (5-HT) is a monoamine neurotransmitter that is synthesized in both the central nervous system and in the chromaffin cells of the gastrointestinal tract. There are numerous 5-HT receptor subtypes: in the brain these receptors occur mainly in the limbic system and the hypothalamus. 5-Hydroxytryptamine modifies mood and behaviour, and a number of different types of drug that increase the effects of 5-HT are used as antidepressants. Drugs that either mimic 5-HT at its receptors or inhibit its reuptake at synapses generally also reduce feeding. The mechanism for this is not clear.

Peripheral signals of appetite

Peripheral signals influencing food intake can be broadly divided into those that cause satiety and those that are secreted in proportion to the amount of fat in the body. The gut and the brain seem to have a close evolutionary relationship. Peptides discovered in the hypothalamus have also been identified in the gut, and gut peptides have been found to be produced by the brain.

Gastric emptying and stretching. Slow gastric emptying increases stomach distension, which activates stretch receptors. The vagus nerve carries afferent signals related to stomach distension to the NTS, facilitating satiety by projections to the appetite-regulating nuclei of the hypothalamus, for example the PVN. Cholecystokinin (CCK) is a potent inhibitor of gastric emptying by vagal afferent-mediated central mechanisms, and this may explain its anorectic actions. Peptide YY and GLP-1 (see below) may also cause anorexia by reducing gastric emptying in this way.

The main neural connection from the gastrointestinal tract to the brain is through the vagal afferent fibres, via the nodose ganglion. Sympathetic afferent fibres in the spinal nerves also carry satiety signals to the brain stem, as previously described.

Hormones. The endocrinological capacity of the gut is diverse, as diffuse populations of endocrine cells are scattered throughout the mucosa. Primary gastrointestinal functions such as motility, secretion and absorption are regulated by gut hormones, which simultaneously provide feedback to the CNS on the availability of nutrients, thereby regulating food intake.

Insulin. Insulin was the first hormonal candidate, related to adipose tissue, to be postulated as the circulating factor that regulates the hypothalamic control of food intake. Insulin was also thought to be important in achieving the usual long-term stability of body weight

and fat mass. The rise in circulating insulin in response to a glucose load is proportional to fat mass. Insulin reaches the CNS via receptor-mediated transport across the BBB and through areas of relative permeability. Central nervous system administration of insulin to rodents causes a reduction in food intake, while brain-specific insulin receptor knockout mice and insulin receptor substrate-2 knockout mice develop obesity. However, when insulin is commenced in patients with type 2 diabetes, weight gain rather than weight loss is observed, possibly as a result of a loss of the anorexigenic effect of hyperglycaemia and the lipogenic actions of insulin.

Cholecystokinin. Cholecystokinin was the first gut hormone described to relay the signal of nutrient intake to the brain, thus leading to the inhibition of further food intake. Cholecystokinin is produced by endocrine cells (I cells) present within the jejunum and duodenum and it is also found in enteric nerves in the ileum and colon. Plasma concentrations of CCK increase in response to the intraluminal presence of the digestion products of protein and fats. Gastric emptying is potently inhibited by CCK through a vagal afferent-mediated central mechanism. In addition to causing satiety via vagally mediated pathways, CCK can cross the BBB and bind directly to specific receptors in the area postrema.

Peptide YY. Peptide YY is secreted from the endocrine L cells of the small and large bowel. It is a 36-amino acid peptide related to NPY. The highest PYY tissue concentrations are in the distal gastrointestinal tract. Postprandial PYY concentrations are proportional to meal energy content. Two isoforms exist, and both PYY_{1-36}} and PYY_{3-36}} may have local effects on gut motility, inhibit secretion of gastric acid and pancreatic enzymes, and inhibit gallbladder emptying. Obese individuals have lower circulating concentrations of PYY.

It is proposed that PYY_{3-36}}, released into the circulation after a meal, inhibits appetite by acting directly on the arcuate nucleus via the Y2R, a presynaptic inhibitory autoreceptor, leading to an inhibition of the NPY neurons and a possible reciprocal stimulation of the POMC neurons. There is a tonic GABA-mediated inhibition of POMC neurons by NPY neurons, and thus decreased GABA-mediated tone as produced by leptin may lead to disinhibition of POMC neurons. Thus, peripheral PYY_{3-36}} reduces the expression of NPY mRNA and increases that of POMC. Exogenous PYY_{3-36}} reduces food intake in rodents, rhesus monkeys and in normal weight and obese humans.

Pancreatic polypeptide (PP). Pancreatic polypeptide (PP) is thought to have arisen by gene duplication of the PYY gene, as PP and PYY are closely related structurally. The PP cells of the pancreatic islets produce PP in response to ingestion of food. This release is in proportion to the energy ingested, and postprandial concentrations remain elevated for up to 6h. Obese subjects have low PP concentrations, while high concentrations of PP have been demonstrated in patients with anorexia nervosa. Peripheral administration of PP to rodents and humans has been shown to reduce food intake.

Leptin. Leptin is a 167 kDa peptide, the product of the *ob* gene, and is produced predominantly by adipose tissue.

Circulating leptin concentrations are directly proportional to adiposity and correlate better with total fat mass than with body weight. Central or peripheral administration of leptin in rodents causes a profound decrease in food intake and weight loss. The *ob/ob* mouse, completely deficient in leptin, is hyperphagic, hyperinsulinaemic and very obese, and the leptin receptor-defective *db/db* mouse has a similar phenotype. Chronic administration of leptin to the *ob/ob* mouse results in sustained reduction in body weight and reduced food intake, but has no effect on the *db/db* mouse.

Leptin deficiency in humans is rare, although it has been suggested that some obese humans have lower leptin concentrations than would be expected. Most obese people have normal leptin genes and have elevated leptin concentrations, reflecting their high adiposity.

Ghrelin. Ghrelin is a 28-amino acid peptide identified in 1999 as the endogenous agonist at the growth hormone secretagogue receptor, and is present in the circulation of healthy individuals. Ghrelin is synthesized in the endocrine cells of the stomach and (in much smaller amounts) in the hypothalamus. Ghrelin has been shown to increase NPY and AgRP in the hypothalamic arcuate nucleus, thus being postulated to increase appetite and act as an initiator of feeding. Plasma concentrations of ghrelin peak before a meal and fall rapidly after nutrient ingestion, supporting the hypothesis that it has a role as a stimulus to eating by signalling pre-meal hunger. A state of negative energy balance increases fasting plasma ghrelin concentrations, while concentrations are reduced by positive energy balance. This suggests that ghrelin is also important in long-term regulation of body weight in normal individuals. Exogenous administration of ghrelin increases appetite and *ad libitum* caloric intake, as well as growth hormone secretion.

Insulin resistance has been postulated to play a role in determining low fasting ghrelin concentrations. Attenuated postprandial reduction of ghrelin concentrations has been shown in the obese, while in normal weight individuals, postprandial suppression of ghrelin secretion is proportional to the calories consumed.

Glucagon-like peptide-1. Processing of proglucagon in the pancreas results in the formation of glucagon, glicentin-related pancreatic polypeptide (GRPP) and major proglucagon fragments. In intestinal L-cells, proglucagon processing results in glicentin, oxyntomodulin (OXM), GLP-1, spacer peptide-2 (SP-2) and glucagon-like peptide-2 (GLP-2). Glucagon-like peptide-1 is co-secreted with PYY in response to the presence of nutrients in the gut.

Glucagon-like peptide-1 is an incretin, stimulating the release of insulin, while also inhibiting the release of glucagon; GLP-1 receptor agonists are used in the treatment of type 2 diabetes. Glucagon-like peptide-1 reduces gastric emptying and inhibits gastric acid secretion, while also regulating food intake. Central administration of GLP-1 in rodents strongly inhibits feeding, and GLP-1 also seems to produce a small, dose-dependent reduction in food intake in humans.

Management of obesity

The objectives of the management of obesity are to reduce body weight and then to maintain a lower body

weight in the longer term. A criterion for success of a reduction of 5–15% in body weight has been proposed. Various approaches can be used to achieve this, alone, or in combination.

Non-surgical options

Dietary approaches offer the obese patient education in the principles of healthy eating and a regimen intended to achieve a deficit in energy balance. Very low calorie diets (400–500 kcal/24h) have been shown to produce greater initial weight loss than low calorie diets (1000–1500 kcal/24h), but results after one year are not significantly different.

It is difficult to achieve weight loss merely through physical exercise. For example, a man of average build who completes a marathon in 4.5 h will only burn about 2000 kcal, approximately the same amount of energy that he would be expected to consume during an entire less active day. He would have to run a marathon every ten days, without an increase in energy intake, to lose a kilogram a month. However, physical activity can contribute towards an energy deficit and is important in increasing cardiorespiratory fitness.

For some obese patients, learned patterns of eating behaviour may be amenable to behaviour modification. Establishment of new patterns of eating and physical activity makes other strategies for weight loss more likely to be successful.

Various drugs have been used for the treatment of obesity. Many have been discontinued due to unacceptable side-effects, including recently, sibutramine, a centrally acting agent affecting reuptake of serotonin, norepinephrine and dopamine; and rimonabant, an inhibitor of cannabinoid receptor-1. Orlistat is a potent inhibitor of pancreatic lipase and, therefore, increases faecal fat loss when taken orally with food. Adverse effects include abdominal distension, flatulence, liquid stools and occasional faecal incontinence. Some of the weight loss in those taking the drug may be a consequence of a conscious decrease in fat intake to avoid these side-effects. Less than 1% of an oral dose is absorbed, so it has minimal interaction with most drugs. However, absorption of fat soluble vitamins may be affected and supplementation may be required. Methylcellulose is available for use as a bulking agent, with the idea that it produces a feeling of satiety, but there is little evidence to support its use. Thyroid hormones have no role in the treatment of obesity, except in patients with biochemically proven hypothyroidism.

Bariatric surgery

The limited long-term success of behavioural and drug therapy for obesity has led to increased use of surgical treatments, with considerable success. Bariatric surgery is considered an option for selected patients with a BMI ≥ 40 kg/m², or those with a BMI > 35 kg/m² with significant comorbidities (e.g. diabetes mellitus), when less invasive methods of weight loss have been unsuccessful and the patient is at high risk of obesity-associated complications.

Possible contraindications to surgery include significant mental health problems, an intercurrent medical

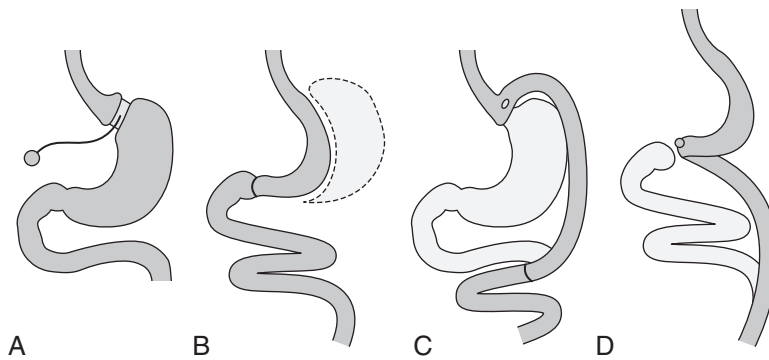


FIGURE 11.1 ■ Bariatric surgery procedures. (A) Adjustable gastric band (AGB). (B) Vertical sleeve gastrectomy (VSG). (C) Roux-en-Y gastric bypass (RYGB). (D) Biliopancreatic diversion with a duodenal switch (BPD-DS).

condition likely to influence surgical risk, a history of poor compliance with medical treatment and pregnancy. Patients should be managed by a multidisciplinary team, including the surgeon, a metabolic physician, dietitian and psychiatrist. Bariatric procedures tend to be described as either restrictive or malabsorptive according to the nature of the surgery involved (see [Figure 11.1](#)). However, these descriptions do not define the mechanism of weight loss.

Restrictive procedures. Laparoscopic gastric banding involves insertion of an inflatable silicon gastric band horizontally around the proximal part of the stomach. The gastric band is connected to a subcutaneous injection port which, when inflated with saline, leads to constriction and the formation of a gastric pouch.

In a laparoscopic sleeve gastrectomy, over three-quarters of the stomach is removed. The gastric remnant is used to fashion a sleeve of stomach, rather than a pouch, which extends from the oesophagus to the duodenum.

Combined restrictive and malabsorptive procedures. Roux-en-Y gastric bypass requires the formation of a small proximal gastric pouch, which empties into a segment of jejunum brought up to the pouch as a Roux-en-Y loop. It results in a degree of malabsorption as a result of bypass of the upper small intestine.

Biliopancreatic diversion with or without duodenal switch involves removal of part of the stomach, together with an alteration in the path of the small intestine.

Most patients report a reduction in appetite after bariatric surgery, an effect more marked after biliopancreatic procedures. Weight loss after bariatric surgery is often reported as the percentage of excess weight loss (%EWL), where excess weight is the total amount above ideal weight. In some series, the %EWL has been reported to be up to 46% after gastric banding, 60% after Roux-en-Y bypass and 64% after biliopancreatic diversion. In addition to weight loss, there is resolution of type 2 diabetes mellitus in many patients and improvement in dyslipidaemia, hypertension and obstructive sleep apnoea. The percentage of patients with resolution of diabetes varies according to the procedure performed, being highest after biliopancreatic

diversion, but it is not directly related to %EWL. This resolution can be very rapid. This suggests that changes in glucose homeostasis are operation-specific and may be determined by intestinal factors. The stomach is the major producer of the orexigenic hormone ghrelin and the functional area of the stomach is reduced in the majority of these procedures. However, postoperative changes in ghrelin are variable and its role in weight loss is unclear. The concentration of the anorectic hormone PYY increases early after malabsorptive procedures, probably due to the re-routing of gut contents, and is thought to be an important contributor to the subsequent weight loss. Incretins, for example GLP-1 and oxyntomodulin, are thought to play a synergistic role in postoperative weight loss.

The mortality rate after bariatric surgery is <1% for younger patients with a BMI <50 kg/m² and average complication rates are <10%. Early complications are related to the surgery itself, for example postoperative bleeding and anastomotic leak. Longer-term problems include erosion from, or displacement of, a gastric band, anastomotic stricture and dumping syndrome owing to the rapid passage of food into the intestine, as well as various metabolic problems, for example metabolic bone disease and kidney stones.

Following bariatric surgery, long-term follow-up is essential. After surgery, patients should be advised to follow a well balanced, calorie-controlled diet. They are usually advised to avoid carbonated drinks, high sugar foods and foods with textures that are difficult to chew such as tough meat and bread. It is also recommended that solid food be chosen for meals, with drinks taken in between. Vitamin and mineral supplementation is recommended after bariatric surgery. After combined restrictive and malabsorptive procedures, but not laparoscopic banding, biochemical assessment of vitamin and trace elements should form part of ongoing review. One suggested scheme for such monitoring is shown in [Table 11.3](#). It is now recognized that a significant proportion of obese patients are deficient in vitamins and trace elements preoperatively. The aetiology includes nutritional deficiency from a diet that, whilst excessive in energy, is likely to have been of poor quality.

TABLE 11.3 Suggested monitoring scheme after bariatric surgery

Roux-en-Y gastric bypass	Biliopancreatic bypass
First year: 3–6-monthly Thereafter: annually	First year: 3-monthly Thereafter: 3–6-monthly depending on results and clinical status
Full blood count	Full blood count
Renal profile	Renal profile
Liver profile	Liver profile
Plasma glucose	Plasma glucose
Lipid profile	Lipid profile
Iron, ferritin, vitamin B ₁₂	Iron, ferritin, vitamin B ₁₂
Vitamin D ₁₂	Fat soluble vitamins (6–12-monthly)
Also consider: PTH, thiamin, folate	A, E, D and prothrombin time Trace elements (annually) Zinc, selenium Metabolic bone evaluation (6–12-monthly) PTH, 24 h urinary calcium Metabolic stone evaluation (12-monthly) 24 h urine calcium, citrate, urate and oxalate

EATING DISORDERS

Eating disorders, which include anorexia nervosa, bulimia nervosa and atypical eating disorders, are a range of syndromes involving physical, psychological and social features. About 1 in 250 females and 1 in 2000 males suffer from anorexia nervosa and about five times as many from bulimia. Atypical eating disorders may be more common, with many patients going untreated.

Anorexia nervosa

Anorexia nervosa is characterized by weight loss (or in children lack of weight gain), leading to a body weight of at least 15% below the normal or expected weight for age and height. The weight loss is the result of patients perceiving themselves to be too fat and so setting a low target body weight that is achieved by food avoidance and, sometimes, obsessive exercise. Secondary to weight loss, patients with anorexia nervosa show evidence of endocrine disturbance involving the hypothalamo–gonadal axis with amenorrhoea in females and decreased libido or impotence in males.

Patients with anorexia nervosa may become severely emaciated and treatment includes nutrition support and psychological therapies. Such patients are at risk of refeeding syndrome (see p. 211), but may also exhibit other biochemical abnormalities prior to reintroduction of appropriate nutrition. Hypokalaemia is frequent in patients with eating disorders; causes include vomiting and loss resulting from misuse of laxatives and diuretics. Hypomagnesaemia is also common and may result from reduced intake and be secondary to laxative and diuretic abuse. Patients who vomit and abuse diuretics are also at risk of hypochlorhaemic metabolic alkalosis. Prolonged use of diuretics may lead to hyponatraemia. Patients who

‘water load’ prior to being weighed at clinic appointments may exhibit a dilutional hyponatraemia. Elevation of liver enzyme activities may be seen in anorexia. This is usually transient and relatively mild, although rarely, severe liver damage occurs; the abnormalities may occur before or during refeeding and suggested mechanisms include liver ischaemia, glutathione depletion and fatty change occurring on refeeding.

At low body weight, there are changes in the hypothalamo–pituitary–gonadal axis. Menstruation ceases and the endocrine changes are of hypogonadotrophic hypogonadism with low concentrations of follicle stimulating hormone, luteinizing hormone and oestradiol. These changes are thought to be related to low concentrations of leptin, which has a role in gonadotrophin releasing hormone secretion. In males, comparable changes occur with low concentrations of testosterone. Plasma concentrations of free thyroxine and free triiodothyronine tend to be low in anorexia, with increased concentrations of reverse triiodothyronine. The thyroid stimulating hormone response to thyroid releasing hormone is abnormal in some patients and it has been suggested that this also is secondary to reduced leptin secretion. In contrast, plasma cortisol concentrations tend to be increased. These and other endocrine changes in patients with anorexia nervosa are summarized in [Box 11.1](#).

BOX 11.1 Endocrine abnormalities found in anorexia nervosa

The hypothalamo–pituitary–adrenal axis

- Increased plasma concentrations of cortisol
- Increased cortisol production rate
- Decreased metabolic clearance of cortisol
- Abnormal dexamethasone suppression test (either failure of suppression or ‘early escape’)
- Increased 24 h urinary excretion of free cortisol

The hypothalamo–pituitary–gonadal axis

- Secondary amenorrhoea
- Low plasma oestradiol concentrations
- Low plasma luteinizing hormone (LH) and follicle stimulating hormone (FSH) concentrations
- Normal LH/FSH response to gonadotrophin releasing hormone (GnRH) following priming

Other hormonal abnormalities

- Exaggerated growth hormone (GH) response to growth hormone releasing hormone (GHRH)
- Reduced GH response to apomorphine; abnormal release of GH in response to thyrotrophin releasing hormone (TRH)
- GH response to a glucose load may show a paradoxical rise
- GH response to insulin hypoglycaemia may be blunted
- Low plasma free T₄ and free T₃ concentrations
- High plasma reverse tri-iodothyronine (rT₃) concentrations
- Delayed thyroid stimulating hormone (TSH) response to TRH
- Subnormal arginine vasopressin (AVP) response to increased plasma sodium concentrations

Bulimia nervosa

Bulimia nervosa is also characterized by a self-perception of being too fat. However, in contrast to anorexia, there are recurrent episodes of overeating. To attempt to counteract the fattening effects of the food consumed, the patient may alternate overeating with periods of starvation, or may induce vomiting and use drugs such as diuretics, laxatives and appetite suppressants to attempt weight reduction.

The endocrine disturbances in bulimia are milder than those in anorexia nervosa. Patients have mild hypercortisolism and basal thyroid hormone concentrations are normal. The secretion of cholecystokinin in response to a meal is impaired in bulimia. About 50% of patients have menstrual abnormalities, which may be one of two types. Patients with low body weight have low plasma oestradiol concentrations and failure of follicular development, while those with normal body weight have normal plasma oestradiol concentrations, but a failure of progesterone secretion in the luteal phase.

Again, the diagnosis is a clinical one, but biochemical tests are important in looking for deficiencies, for example of potassium. Interestingly, in view of the putative role of serotonin in feeding, monoamine oxidase inhibitors, tricyclic antidepressants and fluoxetine have been shown to have beneficial effects in bulimia. However, the effects are smaller than those produced by non-pharmacological therapies and are not maintained.

DIET IN THE AETIOLOGY OF DISEASE

Diet is known to modulate factors that are important in the aetiology of cardiovascular disease (CVD), and dietary change is an important element of CVD risk management. This topic is discussed in detail in Chapter 38. Diet is also important in the aetiology of some gastrointestinal diseases (e.g. coeliac disease), and this is discussed in Chapter 12. Other conditions in which diet plays a causative role include dental caries and some cancers.

Dental caries

The most common reasons for the loss of teeth are dental caries in children and periodontal disease in adults. Periodontal disease has only a minor association with nutrition, but there is a strong relationship between the formation of dental caries and diet.

The first stage of cariogenesis is the formation of plaque. Plaque begins to develop on the surface of a clean tooth by the adsorption of salivary proteins and glycoproteins to form the pellicle. Bacteria (e.g. *Streptococcus mutans*) then become attached to the pellicle and begin to reproduce, eventually forming a continuous layer. A polysaccharide coating then develops over them, formed from the action of bacterial enzymes on dietary sugars. This coating helps to maintain plaque integrity, but also acts as a source of energy for the bacteria during the gaps between the host's meals. Dietary sugars diffuse into the plaque and are fermented by the plaque bacteria into mainly lactic acid, which then dissolves the mineral phase of tooth enamel, causing caries. Both the amount and frequency of sugars

in the diet are positively correlated with caries formation. Some foods (e.g. milk, cheese and the sugar substitute xylitol) protect against dental caries, as does increasing salivary flow with sugar-free chewing gum.

Dietary fluoride, for example in fluoridated drinking water, is important in preventing the development of dental caries. A concentration of 1 mg/L increases the resistance of the enamel, particularly if the exposure occurs while it is being laid down prior to the tooth erupting, and reduces tooth decay in children by about 50%.

TABLE 11.4 Some examples of associations between dietary factors and cancer

Site of cancer	Associated dietary factors
Mouth, pharynx and larynx	Alcohol, very hot drinks, Chinese-style salted fish
Oesophagus	Alcohol, very hot drinks, nitrosamines (in preserved foods)
Stomach	Smoked fish, pickled foods, cured meats, salt; nitrosamines formed in the stomach from nitrites in the diet
Large intestine	High fat, high protein diet, low fibre intake
Liver	Aflatoxin-contaminated foods (aflatoxins are a product of the fungus <i>Aspergillus flavus</i>), alcohol
Breast	High fat diet, alcohol

Cancer

Many associations have been noted between specific dietary components and various cancers (see Table 11.4). The overall risk of cancer appears to be highest in populations that consume diets high in fat and energy, in which there is a high prevalence of obesity and in which alcohol intake is high. Diets that are high in 'fibre', particularly those with a high fruit and vegetable content, appear to offer some protection from certain cancers. However, research in this area is difficult because, although dietary factors may predispose to or protect against cancer, the effect may be small compared with other known carcinogens (e.g. cigarette smoke), there may be interactions between different dietary constituents and methods for the measurement of dietary intakes are inherently inaccurate. Furthermore, altering the amount of any one component of the diet necessarily alters the relative amounts of the others.

THERAPEUTIC DIETS, DIETARY SUPPLEMENTS AND NUTRACEUTICALS

Details of therapeutic diets can be found in textbooks of dietetics, but the main features of some of the more common ones are summarized in Table 11.5.

Apart from the well-established benefits of therapeutic diets, claims are also made that some dietary constituents have health benefits beyond their purely nutritional role. Functional foods are defined as foods that provide a health benefit beyond basic nutrition, when consumed in normal amounts as part of a varied diet. Nutraceuticals are dietary supplements that provide a concentrated form of an agent from a food, in a dose exceeding that obtainable from a

TABLE 11.5 The main features of some therapeutic diets

Disease	Main features of dietetic management
Diabetes mellitus	Reduce risk of microvascular disease by achieving near normal glycaemia, while avoiding risk of hypoglycaemia Reduce risk of macrovascular disease, including management of body weight, dyslipidaemia and hypertension 45–60% dietary energy from carbohydrate, especially from complex carbohydrates or those with low glycaemic index May have up to 10% energy from sucrose, provided it is eaten in the context of a healthy diet and spread throughout the day <10% energy from saturated and <i>trans</i> fatty acids <10% energy from polyunsaturated fatty acids, including ω 3-polyunsaturated fatty acids from oily fish 10–20% energy from monounsaturated fatty acids Protein not more than 1 g/kg/day, i.e. not a high intake
Obesity	Aim for 10% weight loss. Further targets can be set if appropriate Reduce dietary energy by reducing fat and sugar intake Encourage a regular meal pattern, including complex carbohydrates at each meal Encourage physical activity
Undernutrition, e.g. caused by dysphagia, cancer, loss of appetite	Increase protein and energy intake by food fortification measures, energy and nutrient-dense foods, and nourishing drinks Prescribe nutritional supplements if measures above fail
Hyperlipidaemia	Reduce saturated fat intake to 10% dietary energy Partially replace saturated fats with monounsaturated fats and complex carbohydrates Increase ω 3-polyunsaturated fatty acid intake At least five portions of fruit and vegetables per day Reduce salt intake Avoid excess alcohol Lose weight if obese
Celiac disease	Eliminate all gluten-containing foods (wheat, barley, rye and possibly oats) from the diet
Allergy	Exclusion diets, e.g. exclusion of wheat, milk, eggs or additives
Renal disease	Normal protein intake, 0.8–1.0 g/kg/day (low protein diets are now rarely used) Maintain energy intake Adjust sodium, potassium and fluid intake depending on whether predominant retention or loss
Inherited metabolic diseases, e.g. phenylketonuria, galactosaemia	Specialist dietary advice required

normal diet. All such products are sold as foods, or food supplements, rather than drugs, and so are not subject to pharmaceutical licensing regulations. There is, therefore, little evidence and there are few clinical trials to enable assessment of their efficacy and formal documentation of side-effects. It has been noted that kava (*Piper methysticum*), which is said to help in anxiety, is probably hepatotoxic and preparations of St John's wort (*Hypericum perforatum*) are known to decrease the efficacy of antiretroviral drugs, ciclosporin, tacrolimus and some antiepilepsy drugs.

A probiotic is a preparation containing defined microorganisms in sufficient numbers to alter the microflora of the host and exert a beneficial health effect. Such preparations are commonly manufactured as a dairy product supplemented with species such as *Bifidobacteria* or *Lactobacilli*. They have been shown to be effective in reducing the duration of antibiotic-associated diarrhoea and alleviating symptoms and promoting digestion in lactose malabsorption.

Other dietary supplements for which there is less good evidence of benefit include coenzyme Q10 (for mitochondrial disorders, heart failure and ischaemia–reperfusion injury), saw palmetto (for benign prostatic hypertrophy) and taurine, carnitine, creatine, chondroitin and glucosamine (for osteoarthritis). The American Association of Clinical Endocrinologists has produced a useful summary of the available evidence (see Further reading, below).

PROVISION OF NUTRITION SUPPORT

Deficiencies of specific vitamins or minerals are treated by appropriate supplementation of individual nutrients, and this could be thought of as a form of nutritional support. However, the term is usually used to mean the provision of balanced mixtures of nutrients to replace all, or a major part, of normal food intake. It is recommended that nutrition support is provided by a multidisciplinary team (e.g. gastroenterologist, pharmacist, clinical biochemist, dietician, nutrition nurse) that assesses patients on an individual basis and makes adjustments to their nutritional therapy as required. Such an approach has been shown to improve clinical outcomes and reduce morbidity and mortality.

Indications for nutrition support

The basic indications for nutrition support are intestinal failure (e.g. after massive bowel resection) or inability to eat (e.g. in neurological diseases that affect swallowing). However, it should be considered in any patient who is malnourished or likely to become so (see Box 11.2).

Simple measures to support nutritional intake include food enrichment, for example by the addition of cream to soup, or cheese to potato. Various proprietary preparations exist: modular supplements, typically in powder

BOX 11.2 Indications for nutrition support

Nutrition support should be considered in people who are malnourished, as defined by any of the following:

- a BMI of $<18.5 \text{ kg/m}^2$
- unintentional weight loss $>10\%$ within the last 3–6 months
- a BMI of $<20 \text{ kg/m}^2$ and unintentional weight loss $>5\%$ within the last 3–6 months.

Nutrition support should be considered in people at risk of malnutrition, as defined by any of the following:

- have eaten little or nothing for more than five days and/or are likely to eat little or nothing for the next five days or longer
- have a poor absorptive capacity, and/or have high nutrient losses and/or have increased nutritional needs from causes such as catabolism.

form, contain just one macronutrient, and a range of fortified puddings provide both protein and energy. Fortified juice-based products and nutritionally complete, milky-type drinks are also available.

For patients whose intake remains poor, despite use of such measures, or in whom their use is inappropriate, nutritional support via the enteral (gut) or parenteral (intravenous) route is required. Whenever possible, the gut, rather than the intravenous route, should be used to deliver nutritional support. There are several reasons for this. There appears to be no metabolic advantage in intravenous feeding and there are some positive advantages to enteral feeding. Nutrients are needed in the gut lumen to maintain the structural and functional integrity of the gastrointestinal tract. Enteral feeding also stimulates gallbladder motility and reduces the risk of cholelithiasis. Only a limited proportion of certain nutrients is actually absorbed from the gut and, in general, the enteral requirements for specific nutrients are better understood than parenteral requirements. Enteral feeding can supply the gut-specific substrates glutamine and short chain fatty acids, which are not usually present in commercial parenteral feeding solutions. Enteral feeding is also less hazardous than use of the intravenous route and, in financial terms, is much the cheaper of the two options.

In some situations, it may be difficult to decide which route is more appropriate. Particularly in postoperative patients, gut function tends to be assessed by clinical indicators of gut motility, for example the presence of bowel sounds and the passage of flatus or faeces. It is well-established that in such patients, intestinal function tends to return before motility. It may be safe, therefore, to commence enteral feeding cautiously before bowel sounds are heard.

Estimates can be made of the basal metabolic rate in order to tailor feeds to individual nutritional needs. However, guidelines produced by the UK National Institute for Health and Care Excellence (NICE) suggest that, for patients who are not severely ill or injured, or at increased risk of refeeding syndrome (see later), feeds should supply the following:

- 25–35 kcal/kg body weight/day total energy (including that derived from protein)
- 0.8–1.5 g protein (or equivalent as amino acids) (0.13–0.24 g nitrogen)/kg per day
- 30–35 mL fluid/kg (with allowance for extra losses, e.g. from drains or fistulae, and extra input, e.g. from intravenous drugs)
- adequate electrolytes, minerals and micronutrients and fibre if appropriate.

Enteral feeding

The route of administration of enteral feeding tends to be determined by the likely duration of feeding, although physical constraints, for example oral surgery, may need to be taken into consideration. For the majority of patients requiring short-term enteral feeding, the most appropriate method of delivery is via a fine-bore nasogastric feeding tube. Placement of the tube is relatively simple, although its correct positioning in the stomach should always be checked before use by ensuring aspiration of acidic gastric contents (using pH paper), lest inadvertent pulmonary (or, rarely, intracranial) misplacement has occurred.

There is a risk of gastric regurgitation and subsequent pulmonary aspiration during nasogastric feeding and, to minimize this, patients should be nursed at 45° wherever possible. For patients at increased risk of aspiration (e.g. those on ventilators or being nursed flat after certain neurosurgical procedures, diabetic patients with neuropathy) insertion of a feeding tube into the duodenum or jejunum may lessen the risk.

For long-term enteral feeding the preferred route is often a tube placed through the abdominal wall directly into the stomach – a gastrostomy. This avoids local problems such as rhinitis and nasal erosions, which can be associated with prolonged use of a nasogastric tube, and is more secure. This type of gastrostomy is usually inserted percutaneously under endoscopic control (PEG).

The exact nature of the enteral feed given is dependent upon the patient's nutritional requirements, the indication for feeding and any comorbidity, for example renal failure. Whilst a normal diet could be liquefied and used for enteral feeding, this is not generally recommended owing to the risk of infection. Instead, proprietary feeds are preferred. These tend to be of three types: polymeric, pre-digested or disease-specific; they are typically nutritionally complete, may contain fibre and generally contain 1–1.5 kcal/mL.

In polymeric feeds, the nitrogen is in the form of whole protein, the carbohydrate source is partially hydrolysed starch and the fat contains long chain triglycerides. In patients with significant malabsorption, a pre-digested feed may be appropriate. In this, the nitrogen is supplied in the form of short peptides and the fat component contains both long and medium chain triglycerides. In severe malabsorption, elemental feeds providing nitrogen as amino acids and carbohydrate as glucose, with very little fat, can be used. Feeds can also be manufactured with their components adjusted to be suitable for the patient's underlying condition, for example low sodium and volume in patients with liver failure.

Parenteral nutrition

In a patient whose gut is not functioning (or is not functioning sufficiently well to sustain adequate nutrition), parenteral feeding has to be used. Parenteral nutrition (PN) solutions are hyperosmolar and tend to be irritant when infused via peripheral cannulae. For short-term feeding, such administration is possible (particularly if hyperosmolarity is reduced by having fat as a major energy source), but infusion into a large vein is preferred. This is usually achieved via a midline, peripherally inserted central line (PICC) or a central venous line. For longer-term feeding, a tunneled line, such as a Hickman, is preferred, to reduce the possibility of infection. Ideally, lines should be used solely for PN. However, if vascular access is compromised, a line with multiple lumens may be provided, but one lumen should be specified and reserved for feeding.

Composition of parenteral nutrition fluids

The carbohydrate component of parenteral nutrition solutions is provided as glucose. Protein is given as a mixture of L-amino acids. There are relatively few data to enable characterization of the optimal quantity of total or individual amino acids. Parenteral glutamine has been shown to be of benefit in certain groups of patients, although the exact situations in which its use is appropriate, and the most appropriate dose, are not yet fully defined. There are, as yet, no definite clinical recommendations regarding supplementation of arginine, branched chain amino acids, cysteine or taurine in parenteral nutrition solutions. Lipid preparations for parenteral use consist of oil stabilized in an emulsion with egg yolk lecithin. The most widely available preparations worldwide are based on soybean oil, but newer products now exist including olive oil-soybean mixtures, fish oil emulsions and multiple lipids (e.g. 'SMOF' a mixture of soy, medium chain triglycerides, olive and fish oil). Fatty acids have a role in the immune response, for example via their incorporation into cell membranes and effects on membrane fluidity and receptor expression. Particularly in patients in whom a robust inflammatory response is required, or in whom there is a particular risk of exacerbation of oxidative stress, evidence is accumulating that the choice of lipid preparation may have an effect on outcome.

Parenteral nutrition should contain appropriate amounts of micronutrients. In the past, trace elements were added based on knowledge of the requirements of well patients eating a normal diet and instances of deficiency and toxicity occurred. Since then, the recommendations for the micronutrient component of commercial preparations used in PN have been regularly reviewed, the most recent modifications suggesting a decrease in the content of copper, manganese and chromium.

Premixed bags of PN solutions, with a variety of amino acid (or nitrogen), energy and electrolyte content, are readily available from commercial suppliers. The bag with contents most suitable for an individual patient's requirements is chosen and variation in the rate of infusion allows some adjustment to the amount of nutrition administered. For patients whose requirements cannot be

accommodated by such premixed bags, individually prescribed nutrition solutions can be prepared under sterile conditions in many hospital pharmacies.

Complications of parenteral nutritional support

Malnutrition is known to be associated with adverse effects that lead to complications and prolonged hospital stay. Nutrition support has been shown to improve quality of life and to reduce complication rates, infection, mortality and length of hospital stay. However, nutrition support, particularly when provided by the parenteral route, is itself associated with complications.

The main non-metabolic complications are related to the placement and presence of the feeding line. Complications associated with central venous catheterization include arterial puncture, air embolism and pneumothorax. Once inserted, good care of the line is essential, using sterile technique to minimize the major risk of catheter-related sepsis. Other possible complications include line blockage, line displacement and thromboembolism.

The aim of monitoring in patients receiving nutrition support is to ensure efficacy and to prevent or identify associated complications. It is important to keep careful records of the amount of nutrition that the patient has actually received and to chart any additional fluid input and losses. Regular weighing and measurement of skin fold thickness may provide an objective way of charting the effectiveness of nutrition support. A protocol for laboratory monitoring of nutritional support has been proposed by NICE (see [Table 11.6](#)). Other aspects to be monitored, particularly in patients on long-term feeding, include the course of the primary disease and the patients' quality of life.

Electrolyte abnormalities are common in patients receiving PN and are most likely to occur soon after initiation of treatment. Abnormalities of sodium, potassium and magnesium are common and factors such as gut losses through vomiting, faecal or stomal loss, renal failure and body water excess, for example from ascites, may be contributory.

Hyperglycaemia is frequent in parenteral feeding, particularly in patients with coexisting sepsis and, if detected, should be controlled with a continuous peripheral insulin infusion: careful control of blood glucose has been shown to improve outcome in critically ill patients. Hypoglycaemia has been documented on abrupt cessation of total parenteral nutrition, but is not commonly encountered with modern feeding regimens.

Hypercalcaemia can occasionally occur in patients receiving parenteral nutrition that includes a vitamin D supplement, and hypocalcaemia can occur if insufficient supplements are given. Hypocalcaemia of unexplained origin should prompt measurement of the plasma magnesium concentration, as hypomagnesaemia may itself cause hypocalcaemia.

The refeeding syndrome comprises a group of abnormalities that may occur when nutrition is provided to those who are severely malnourished. The most well recognized feature of the syndrome is hypophosphataemia, but other electrolyte abnormalities can also occur together with cardiac, pulmonary, neuromuscular and

TABLE 11.6 Protocol for laboratory monitoring of nutrition support

Parameter	Frequency
Full blood count	Baseline 1–2 times a week until stable Then weekly
Renal profile	Baseline Daily until stable Then 1–2 times a week
Liver profile (including prothrombin time)	Baseline Twice weekly until stable Then weekly
Glucose	Baseline 1–2 times a day (more if needed) until stable Then weekly
Calcium, albumin	Baseline Then weekly
Magnesium, phosphate	Baseline Daily if at risk of refeeding syndrome Three times a week until stable Then weekly
Iron, ferritin folate, vitamin B ₁₂	Baseline Every 3–6 months
CRP ²	Baseline As required to assess presence of an acute phase response and assist interpretation of protein, trace element and vitamin results
Zinc, copper	Baseline Then every 2–4 weeks depending on result
Selenium ^a	Baseline if risk of depletion Further testing dependent on baseline
Manganese ^b Vitamin D ^b	Every 3–6 months if on home PN 6 monthly if on long-term support
Bone densitometry ^b	On starting home PN Then every two years

^aNeeded primarily for patients having PN in the community.

^bRarely needed for patients having enteral feeding. PN, parenteral nutrition.

haematological complications. It most commonly occurs in association with parenteral nutrition, although it has been documented as a complication of enteral nutrition. It should be anticipated in those patients at increased risk (see Box 11.3) in whom feeding should be commenced at a low rate and increased gradually.

The aetiology of the hypophosphataemia is probably multifactorial. One important factor is that, during prolonged starvation, energy production is derived principally from fatty acid oxidation rather than from glucose; however, on refeeding and reintroduction of carbohydrate, there is a rapid return to the use of glucose as the predominate substrate with a high requirement for phosphate to synthesize the phosphorylated intermediates and products of glycolysis (e.g. ATP and 2,3-DPG). This leads to hypophosphataemia. There is an intracellular shift of potassium and magnesium as a result of increased insulin release, which leads to hypokalaemia and hypomagnesaemia. Sodium and water retention may occur: an anti-natriuretic action of insulin has been suggested as a possible mechanism.

BOX 11.3**Criteria for identifying people at high risk of developing refeeding problems**

The patient has one or more of the following:

- BMI <16 kg/m²
- unintentional weight loss >15% within the last 3–6 months
- little or no nutritional intake for more than ten days
- low plasma concentrations of potassium, phosphate or magnesium prior to feeding.

or

The patient has two or more of the following:

- BMI <18.5 kg/m²
- unintentional weight loss >10% within the last 3–6 months
- little or no nutritional intake for more than five days
- a history of alcohol abuse or treatment with drugs including insulin, chemotherapy, antacids or diuretics.

Thiamin deficiency is common in severely malnourished individuals. Thiamin is a cofactor in glycolysis and, when carbohydrate intake is restored after caloric deprivation, demands are increased. In thiamin deficiency, pyruvate is converted to lactate, which may lead to lactic acidosis. Precipitation of the complications of thiamin deficiency should be prevented in those at risk of refeeding syndrome by administration of thiamin (initially in amounts in excess of normal requirements) before commencing, and during, nutrition support.

The clinical features of refeeding syndrome are related to the electrolyte abnormalities and vitamin deficiency.

Mild elevation of plasma liver enzyme activities is common after several weeks of parenteral feeding. Liver histology reveals hepatic steatosis with accumulation of macro- and microvesicular fat. However, the histological abnormalities are more closely correlated with the presence of intra-abdominal sepsis, renal failure and pre-existing liver disease than with the duration of PN.

Liver disease is a significant complication of long-term PN, estimated to occur in 15–60% of patients. In adults, steatosis occurs, progressing to intrahepatic cholestasis and then cirrhosis. The incidence of liver disease is higher in children, in whom cholestasis occurs earlier and can be rapidly progressive. Aetiological factors include the nature of the primary disease and residual intestinal anatomy, the presence of sepsis, lack of enteral nutrition and possible nutrient deficiencies or toxicity. Management of PN-induced liver disease includes detection and treatment of sepsis and optimization of the feeding regimen, perhaps with consideration of reduction in energy content. Ursodeoxycholic acid has been used with some benefit, particularly in neonates; rarely, liver and small intestinal transplantation may be required.

The prevalence of metabolic bone disease, characterized by reduced bone mineral density (BMD), is high in patients on long-term PN. Possible contributing factors include abnormalities (insufficiency or excess) of vitamin D, calcium, vitamin K, copper and aluminium. In addition, inflammatory bowel disease, which is an aetiological

factor in a number of the patients, is, itself associated with reduced BMD through the action of bone resorbing cytokines and corticosteroid use. The use of bisphosphonates may be associated with an increase in BMD (as assessed by dual-energy X-ray absorptiometry, DEXA) in patients on PN, but their effect on fracture prevention is unclear.

Short bowel syndrome

For the majority of patients, parenteral nutrition is required for a short period during intercurrent illness. For others, it is a long-term treatment and they incorporate its administration into their daily routine at home and are otherwise well. The majority of these patients will have a short bowel. A short bowel is defined as one that is of insufficient length to enable adequate absorption, so that supplementation of macronutrients, and/or water, and/or electrolytes is essential. This is likely to occur in patients left with <200 cm of small bowel. The most common causes of short bowel syndrome are mesenteric artery thrombosis, Crohn disease and radiation damage. The majority of patients with a short bowel can be divided into two groups: those with a short bowel and a jejunocolic anastomosis, and those with a short bowel and a high output stoma. To facilitate management, the latter group can be divided into net 'absorbers' and net 'secretors'. The absorbers generally have more than ~100 cm of jejunum remaining. They are able to absorb more sodium and water from the diet than they take orally and have low stomal output. Net secretors have very little residual jejunum; stomal output is high (4–8 L/24h) and losses of sodium and water exceed oral input.

Management of a patient with a high output jejunostomy includes oral fluid restriction to reduce stomal output. Any fluid given should contain sodium in a concentration of at least 100 mmol/L, together with glucose, to minimize passive sodium loss and facilitate glucose-coupled absorption of sodium. Antimotility agents (e.g. loperamide and codeine) and antisecretory drugs (e.g. H₂ antagonists, proton-pump inhibitors and somatostatin) may be prescribed to further reduce stomal output.

For some patients with short bowel syndrome, small bowel transplantation is an alternative to long-term parenteral nutrition. Indications for intestinal transplantation include PN-related liver disease, central venous catheter related thrombosis and frequent central line sepsis. Patients who experience frequent episodes of dehydration, despite intravenous fluids may also be eligible. Survival rates after transplantation have shown steady improvement and are currently about 80% at one year and 43% at ten years.

CONCLUSION

Nutritional disorders include a wide spectrum of conditions, including generalized undernutrition, overnutrition leading to obesity, the eating disorders and diseases where nutrition has a role in the aetiology. Globally, both undernutrition and obesity are important public health problems. The treatment of undernutrition is often complicated by factors such as war, famine and infectious diseases. Obesity remains difficult to treat once present, although the advances in the understanding of the physiology of feeding discussed in this chapter are leading to new pharmacological and surgical interventions.

Nutritional interventions in the treatment of diseases may involve the use of therapeutic diets, the administration of dietary supplements or the provision of nutritional support, either enteral or parenteral. Nutritional support is best delivered by a multidisciplinary team. Further research is required to establish the efficacy (or otherwise) of most dietary supplements.

ACKNOWLEDGEMENT

We would like to thank Stephen K. Bangert and Carel W. Le Roux who wrote the chapter in the previous edition of this book.

Further reading

- American Association of Clinical Endocrinologists Nutrition Guidelines Taskforce. Medical guidelines for the clinical use of dietary supplements and nutraceuticals. *Endocr Pract* 2003;9:417–70.
A useful summary of the evidence available for some nutraceuticals.
- American Association of Clinical Endocrinologists, The Obesity Society and American Society for Metabolic & Bariatric Surgery Medical Guidelines for Clinical Practice for the Perioperative Nutritional, Metabolic and Nonsurgical Support of the Bariatric Surgery Patient. *Surg Obes Relat Dis* 2008;S109–84.
This guidance covers assessment of preoperative assessment of bariatric surgery patients and aspects of postoperative nutrition, vitamin supplementation and laboratory monitoring.
- Geissler C, Powers H, editors. *Human nutrition*. 12th ed. Edinburgh: Churchill Livingstone; 2011.
The twelfth edition of the classic text Human Nutrition and Dietetics, previously edited by Garrow, James and Ralph.
- Hing KN, Ang YS. Overview of bariatric surgery for the physician. *Clin Med* 2012;12:435–40.
- National Collaborating Centre for Acute Care. *Nutrition support in adults. Oral nutrition support, enteral tube feeding and parenteral nutrition*. London: National Collaborating Centre for Acute Care; 2006.
- The NICE guidelines on nutrition support in adults: <http://www.nice.org.uk/guidance/QS24>.
- Woods SC, D'Aleesio DA. Central control of body weight and appetite. *J Clin Endocrinol Metab* 2008;93:537–50.
A review of the mechanisms involved in appetite and satiety.

Clinical biochemistry of the gastrointestinal tract

Ingvar T. Bjarnason • Roy A. Sherwood

CHAPTER OUTLINE

INTRODUCTION 214

MOUTH AND OESOPHAGUS 215

STOMACH 215

Helicobacter pylori 215

Diagnosis of *H. pylori* infection 216

Gastric acid secretion 216

Gastrin 216

Intrinsic factor 217

PANCREAS 217

Pancreatic function tests 217

SMALL BOWEL BACTERIAL OVERGROWTH 218

The normal intestinal microflora 218

Definition, causes and symptoms of small bowel bacterial overgrowth 219

Diagnosis of small bowel bacterial overgrowth 219

MALDIGESTION AND MALABSORPTION 219

Clinical features 219

Carbohydrate absorption 220

Protein absorption 222

Fat absorption 224

INTESTINAL PERMEABILITY 225

FAECAL TESTS OF INTESTINAL INFLAMMATION 225

Calprotectin 225

Calprotectin in disease 225

Neuroendocrine tumours of the gastrointestinal tract and pancreas (NETs) 226

THE ACUTE ABDOMEN 227

Introduction 227

Acute pancreatitis 228

Ectopic pregnancy 229

Acute porphyria 230

INTRODUCTION

Gastrointestinal complaints continue to account for a sizeable proportion of medical consultations. This is best illustrated by the fact that irritable bowel syndrome (IBS), the most commonly encountered gastrointestinal disorder, is estimated to have a prevalence of 20% in the UK. About 30% of these people seek medical attention, which in turn accounts for as many as 20–50% of patients attending general gastroenterology outpatient clinics in the UK. Furthermore, the prevalence of reflux oesophagitis and non-specific dyspepsia is increasing, and with the advent of safe and effective drugs to inhibit gastric acid secretion, it is estimated that the cost of treatment for these disorders may grow to account for as much as 20% of drug expenditure in primary care.

The development of endoscopic techniques has revolutionized the investigation of gastrointestinal disorders. The combination of upper gastrointestinal endoscopy, colonoscopy and optic and/or wireless capsule enteroscopy now allows visualization of the whole of the gastrointestinal tract and biopsies can be obtained from all parts of the bowel. This has resulted in loss of demand

for many of the classic biochemical investigations (such as gastric acid secretion) and, as some of these methods are now obsolete, we have omitted them from this chapter. Endoscopy and biochemical investigations, however, provide different kinds of information that are, in many respects, complementary. Endoscopy provides a static morphological picture that by itself, or with biopsy, has the potential to provide a diagnosis that automatically translates to treatment. Biochemical methods, on the other hand, provide a wider range of information. These include the following.

- **Providing a diagnosis.** This is a frequently desired purpose of clinical biochemistry, but is rarely achieved. Nevertheless, the indirect documentation of the presence of *Helicobacter pylori* is often the only test required before treatment, and non-invasive tests for small bowel bacterial overgrowth may be the only tests required in order to treat diabetic patients with diarrhoea.
- **Clinical screening tests prior to invasive investigation.** Examples of these kinds of investigations are serum transglutaminase antibody measurements and tests of intestinal permeability, which, if normal, may

avoid the need for more invasive tests and, if positive, are an indication for jejunal or duodenal biopsy.

- **The detection and assessment of the severity of intestinal dysfunction.** These tests are called upon in order to provide an explanation for clinical signs (such as weight loss), for monitoring response to therapy and for confirmation of diagnosis (e.g. measurement of intestinal permeability following gluten withdrawal and challenge in coeliac disease, although this is rarely necessary in practice).
- **Providing prognostic information.** Biochemical tests are uniquely suited to assess functional changes that may herald a drastic change in the activity of the disease. This is best exemplified by increased concentrations of inflammatory markers in the faeces of patients with clinically quiescent inflammatory bowel disease (IBD, i.e. ulcerative colitis and Crohn disease), as these predict imminent clinical relapse.
- **Investigating the impact of non-intestinal factors on intestinal function, whether biochemical or physiological.** These may be exogenous, for example related to radiotherapy, drugs, alcohol, dietary or environmental factors, or endogenous, for example due to malnutrition, reduced blood flow, anaemia etc. Non-steroidal anti-inflammatory drug (NSAID)-induced enteropathy is an example of gastrointestinal disease caused by an exogenous factor.

Our intention in this chapter is to review new and established laboratory-based investigations of the gastrointestinal tract that are useful clinically or for research purposes. We will not always provide the reference ranges for test results, as these may differ depending on environmental, geographical and racial factors, as well as differences in methodology.

MOUTH AND OESOPHAGUS

Clinical biochemistry has not had a major impact in the investigation of oral or oesophageal diseases. The reason is obvious in that the main functions of these organs are physical, namely grinding of food and transport to the stomach. Nevertheless, the buccal mucosa is a common site for obtaining material for genetic analysis and saliva can be assayed for a number of antibodies (although this is rarely done in clinical practice). Although the salivary glands and parotid secretions contain amylase (and initiate digestion of complex carbohydrates), peptides and growth factors (that may confer a degree of protection to the stomach and accelerate healing of gastric lesions), measurements of these are not helpful clinically and are used only in research.

The common oesophageal disturbances such as reflux and dysmotility (oesophageal spasms, neuromuscular incoordination, achalasia etc.) are investigated and diagnosed by imaging techniques (endoscopy, radiology), high-resolution manometry and pH recording, rather than by biochemical methods. Oesophageal cancer, infections (cytomegalovirus, *Candida albicans* and herpes simplex) and drug- and radiation-induced diseases are similarly diagnosed largely from the clinical history and by endoscopy and biopsy.

STOMACH

The stomach acts as a reservoir for ingested food, where it is mixed with acid, mucus and pepsin and then released at a controlled rate into the duodenum. The mucosal surface of the stomach is lined with mucus-secreting columnar epithelial cells, interrupted by gastric pits containing acid-secreting parietal (or oxyntic) cells and pepsinogen-secreting chief cells. The secreted hydrochloric acid kills many ingested bacteria, provides the acid pH necessary for pepsin (from pepsinogen) to begin digesting proteins and stimulates bile secretion. Mucus secretion is necessary to protect the mucosal surface of the stomach wall from its acidic contents. The stomach also secretes intrinsic factor, which binds vitamin B₁₂ and allows it to be absorbed in the terminal ileum. The muscular contractions of the stomach wall have the mechanical effect of macerating food.

Helicobacter pylori

Helicobacter pylori is now accepted to be the main cause of gastric and duodenal ulcers; other causes include NSAIDs and, very rarely, the Zollinger–Ellison syndrome. *Helicobacter pylori* may play an important role in gastric cancers (adenocarcinomas and mucosal-associated lymphoid tissue lymphomas) and is undoubtedly the main cause of chronic gastritis, though in most patients this is asymptomatic. Many experts also believe *H. pylori* to be a major cause of non ulcer-related dyspepsia, and recent recommendations suggest that anyone who so wishes should be tested for *H. pylori* and treated if found to be positive.

The mode of transmission of *H. pylori* is uncertain. It is thought to be an infection acquired in childhood, presumably by the faeco–oral route, which would explain the higher prevalence in developing countries and the progressive decline in prevalence in developed countries, with better hygiene, preservation of food and smaller family sizes. Whatever the mode of transmission, *H. pylori* elicits an inflammatory response that is usually asymptomatic. This takes the form of chronic gastritis, with an acute inflammatory cell infiltrate of variable severity. Nine out of ten people infected by *H. pylori* do not develop ulcers. The variable development of clinically significant disease relates to the site of infection, virulence factors (e.g. phospholipases, vacuolating cytotoxins (VAC), CagA protein) and poorly defined host factors that include blood flow, mucus secretion and pepsinogen stimulation. There are three recognized patterns of infection by *H. pylori* in the stomach. The most common type is that of low-grade inflammation of the mid-body of the stomach. This occurs in people with a high threshold for immune response and *H. pylori* that has a low expression of CagA and VAC. Gastric acid secretion is reduced and there are usually no significant clinical consequences. If the microbe is predominantly in the antrum, the inflammation leads to dysfunction of the antral G-cells that become hyper-reactive and secrete disproportionate amounts of gastrin in response to food and gastric distension. Gastrin has a trophic effect on the gastric parietal cells, increasing their number and stimulating them directly, via cell surface receptors, to produce more acid, resulting in the hyper-acidic state characteristic of patients with duodenal ulceration.

The third pattern, a pangastric infection with *H. pylori*, is characteristic of those who develop gastric ulcers and those at risk of developing gastric cancer. These individuals tend to have normal or reduced gastric acid secretion. In theory, it would be possible to perform an endoscopy on every patient with indigestion and document the pattern of infection. However, this would be prohibitively expensive owing to the large number of symptomatic patients. A less expensive option would be to assess a panel of hormones that are secreted from defined anatomical locations in the stomach (for instance pepsinogen 1 and 2 and gastrin-17), but further work is required before this approach becomes widely accepted. A more pragmatic approach has, therefore, been adopted in adult patients under 50–55 years of age, namely to test and treat for *H. pylori* infection, irrespective of the precise diagnosis (ulcer, gastritis etc.), unless 'red flag' indicators suggestive of malignancy (e.g. weight loss, dysphagia) are present.

Having established the presence of *H. pylori* (see below), there are a number of eradication regimens available, and the eradication failure rate, after trying at least three different regimens, should be <5–10%. If *H. pylori* is successfully eradicated, reinfection rates are extremely low (<1%).

Diagnosis of *H. pylori* infection

There are various methods for the diagnosis of *H. pylori* infection. At endoscopy, it is possible to take biopsies from which the organism can be visualized histologically or cultured, the latter being performed only if there are repeated treatment failures. Alternatively, the biopsy can be tested directly using commercially available kits incorporating a gel containing urea and an indicator that changes colour at an alkaline pH. In the presence of *H. pylori* (which contains urease), the urea is broken down to carbon dioxide and ammonia. The latter increases the pH of the gel and a colour change takes place. Less invasive methods involve measurement of *H. pylori*-specific antibodies (IgG or IgA). These are eminently suitable for screening (sensitivity 92%, specificity 83%), but cannot be used to document the success of treatment as high antibody titres can persist despite successful eradication.

The *H. pylori* breath test is currently the most widely used method for non-invasive diagnosis. It uses the same principle as the biopsy-based test and comes at a fraction of the cost of endoscopy and biopsy. Patients are given isotopically labelled urea (¹³C or ¹⁴C) to drink. If there is no urease present, the urea is absorbed intact and excreted in the urine. If *H. pylori* is present, labelled carbon dioxide is absorbed into the circulation and exhaled in the breath. Breath samples are obtained before, and 45–60 min after, drinking the labelled urea. The detection of labelled carbon is most commonly performed by mass spectrophotometry. The breath test is widely used and can also be used to assess the success of the treatment (sensitivity 95%, specificity 96%). Lastly, there are commercial kits that use the polymerase chain reaction to amplify nuclear sequences specific for *H. pylori* from saliva or faeces (sensitivity 95%, specificity 94%). The stool tests are increasingly being used for the detection

and confirmation of successful eradication of *H. pylori* rather than the breath test.

Gastric acid secretion

Before the discovery of the role of *H. pylori* in peptic ulcer disease, it was common practice to investigate gastric acid secretion in patients with duodenal ulcers, who tend to be hyper-secretors, and those with gastric ulcers, who tend to have normal or low secretion rates. Gastric cancers are often associated with hypochlorhydria, and achlorhydria is common in pernicious anaemia and in gastric cancer.

Acid secretion tests are now only rarely used in the research setting and are no longer available in the vast majority of clinical biochemistry departments: they have become obsolete as the test results do not alter clinical practice or management.

Gastrin

In <0.5% of patients with gastroduodenal ulcers, the cause is unregulated gastrin release from an endocrine tumour termed a gastrinoma, which can lead to the Zollinger–Ellison syndrome characterized by multiple gastroduodenal ulcers. The persistently high plasma gastrin concentrations not only lead to marked hypersecretion of gastric acid, but also, because the hormone is trophic for parietal cells, to increased parietal cell mass. Patients may present with symptoms that are indistinguishable from *H. pylori*-associated peptic ulcer disease. However, co-existing diarrhoea (due to acidic inactivation of pancreatic enzymes), multiple ulcers involving the second part of the duodenum, recurrent ulceration and ulcers refractory to conventional treatment should always arouse suspicion. Gastrinomas are either sporadic (the more common form) or associated with multiple endocrine neoplasia type 1 (MEN 1, Wermer syndrome), a syndrome characterized by the presence of two or more of pituitary, pancreatic islet and parathyroid tumours. Multiple endocrine neoplasia type 1 is also associated with an increased prevalence of carcinoid, adrenocortical and thyroid tumours.

Gastrinomas are commonly situated in the pancreas, but are increasingly recorded arising from the duodenum, stomach and bones. Some 60% are clearly malignant, with multiple metastases at diagnosis. Given a suspicion of the Zollinger–Ellison syndrome, the first step is to measure fasting serum gastrin, as virtually all cases are associated with high concentrations. The differential diagnosis of a mildly elevated gastrin concentration includes hypochlorhydria, long-term proton pump inhibitor drug therapy, pernicious anaemia and antral G-cell hyperplasia. Given a strong suggestion of a gastrinoma, the next step is to localize the tumour, which is best done at a specialist centre. Techniques for localization of the tumour include endoscopic ultrasound, octreotide scanning, magnetic resonance imaging (MRI), positron emission tomography (PET) and computerized tomography (CT) (the latter being particularly useful to detect metastases). Definitive diagnosis is histological. The treatment of gastrinomas usually involves a combination of surgery, chemotherapy and acid suppression for symptomatic relief.

Intrinsic factor

Intrinsic factor is a glycoprotein principally secreted by the parietal cells of the stomach. Its secretion is governed by the same biochemical processes that regulate acid secretion and its action is to assist in the absorption of vitamin B₁₂. Vitamin B₁₂ is released from dietary proteins by the action of pepsin and then combines with R-binders (haptocorrins), glycoproteins secreted by the stomach, that assist in its protection against acid degradation. It is subsequently cleaved from R-binder complexes by pancreatic proteolysis in the duodenum and then binds to intrinsic factor to form a resistant complex that is taken up by ileal cell receptors (cubulin).

The commonest cause of intrinsic factor deficiency is autoimmunity; the reason for requesting intrinsic factor antibody measurement is usually the discovery of a low plasma vitamin B₁₂ concentration – a cause of macrocytic anaemia and various neurological disorders. The diagnosis of pernicious anaemia is based on the finding of a low plasma B₁₂ concentration together with the presence of antiparietal cell antibodies and/or intrinsic factor antibodies (see Chapter 27). Current automated vitamin B₁₂ assays suffer, in differing amounts, from interference from intrinsic factor antibodies; alternative strategies for assessment of B₁₂ status include measurement of serum homocysteine, methylmalonic acid and holotranscobalamin concentrations. Other causes of intrinsic factor deficiency leading to low plasma B₁₂ include gastrectomy, achlorhydria and congenital absence of intrinsic factor, which is rare. Vitamin B₁₂ deficiency can also occur because of deficient intake (e.g. in vegans, starvation or reduced food intake of any cause) or in small bowel disease involving the terminal ileum. Gastric biopsies showing histopathological features of atrophic gastritis may be helpful and there are some indications that patients should undergo surveillance endoscopies once a diagnosis is made because of an increased incidence of gastric cancer. In patients in whom small bowel disease is suspected, the choice is that of wireless capsule enteroscopy or CT/MRI enteroclysis (specialized techniques in which a contrast medium is infused into the small intestine) for diagnosis. Previously, on finding a low vitamin B₁₂ concentration, it was customary to request a Schilling test, which had the potential to help in the differential diagnosis, but this test has become obsolete as imaging techniques have improved. Although proton pump inhibitors induce a state of gastric acid hyposecretion and are widely used, it is exceptionally rare to see vitamin B₁₂ deficiency in these patients.

PANCREAS

The exocrine function of the pancreas includes the production of bicarbonate and enzymes, including amylase, lipase, trypsin, chymotrypsin, esterases and carboxypeptidases. The differential diagnosis of exocrine pancreatic disease in neonates and children is predominantly between cystic fibrosis and pancreatic acinar cell aplasia (Shwachman–Diamond syndrome). The major diseases affecting the pancreas in adults are acute pancreatitis,

chronic pancreatitis leading to pancreatic insufficiency and carcinoma of the pancreas. Chronic pancreatitis results in progressive loss of both islet cells and acinar tissue. Presentation is typically with recurrent upper abdominal pain radiating to the back, although malabsorption, for example steatorrhoea, may be the presenting feature. Approximately 90% of the pancreatic acinar tissue must be lost before features of malabsorption become apparent and clinically significant reduction in endocrine function generally occurs late in the disease process.

Pancreatic function tests

Tests of pancreatic function are usually divided into the direct (invasive) tests and indirect tests on blood, urine or faecal samples.

Direct or invasive function tests

The ‘gold standard’ test of pancreatic function is the secretin–pancreozymin test. This test assesses exocrine function by measuring bicarbonate and pancreatic enzyme secretion (amylase and trypsin) in aspirates from a tube sited in the duodenum, usually under fluoroscopic control. Secretin is given to induce fluid secretion, while pancreozymin or its analogue caerulein (secretin–caerulein test), is given to induce enzyme production. This test requires meticulous attention to technique in positioning the tube correctly and maintaining its patency, and is uncomfortable for the patient. This test is now rarely used in routine practice. However, it remains the test that has the highest sensitivity and specificity for the differential diagnosis of pancreatic insufficiency.

Non-invasive pancreatic function testing

Serum enzymes. The measurement of pancreatic enzymes in serum is standard practice in acute pancreatitis, but is seldom useful in the investigation of chronic pancreatitis.

Amylase is the most commonly measured enzyme owing to the availability of cheap, easily automated methods. A disadvantage is the lack of specificity for the pancreas, as amylase in the circulation is derived from both pancreatic and non-pancreatic (mostly salivary) sources in approximately equal amounts. Measurement of the specific pancreatic amylase can be achieved using immunosubtraction techniques. Consideration of the ethnicity of the patient is important when interpreting amylase results, as subjects of African origin have a higher reference range for non-pancreatic amylase. The mechanism for this difference is unknown. An increase in both the pancreatic and non-pancreatic fractions may indicate the presence of macroamylasaemia – immunoglobulin–amylase complexes that have a prolonged half-life in the circulation owing to a reduction in clearance. Non-pancreatic causes of an increased serum amylase are given in Table 12.1. Lipase has greater specificity for pancreatic disease than amylase, but methods available tend to be more complex and expensive. The majority of the lipase in blood is the pancreatic form, although a sublingual lipase is also present. Lipase is not affected by ethnicity. Trypsin is 100% specific to

TABLE 12.1 Non-pancreatic causes of an elevated plasma amylase activity

Condition	Reason for hyperamylasaemia
Perforated peptic ulcer	Increased entry of pancreatic enzymes into the circulation
Small bowel obstruction or perforation	Increased entry of pancreatic enzymes into the circulation
Ruptured ectopic (tubal) pregnancy and salpingitis	Release of Fallopian tube amylase
Salivary gland inflammation, e.g. with calculi or mumps	Release of salivary gland amylase
Opiate administration	Contraction of the sphincter of Oddi
Renal failure	Impaired renal clearance of amylase
Macroamylasaemia	Amylase becomes combined with a plasma protein (in some cases, an immunoglobulin), and the increased size of the complex results in impaired renal clearance. Thus, hyperamylasaemia may be detected in the absence of disease
Certain tumours of the lung and ovary	Production of amylase by the tumour
Diabetic ketoacidosis	Impaired renal clearance
Subjects of African origin	Higher circulating amylase of unknown cause

the pancreas and would therefore, theoretically, be the best of the three enzymes to measure. However, there are physiological obstacles that mean that trypsin itself is seldom measured in blood samples. Trypsin is produced and stored in the pancreas as its inactive zymogen form (trypsinogen), and is activated after secretion into the intestinal tract. Active trypsin entering the circulation is bound immediately to the protease inhibitors α_2 -macroglobulin and α_1 -antitrypsin. Once bound, trypsin is not measurable using standard techniques, but any trypsinogen entering the circulation can be measured as 'immunoreactive trypsinogen' (IRT). Immunoreactive trypsinogen can be used as a first-line test for screening for cystic fibrosis using the blood spots taken on Guthrie cards at 7–10 days of age. A national programme of cystic fibrosis screening started in 2006–2007 in the UK; IRT is used for initial screening with subsequent genetic testing for confirmation.

Faecal tests. The measurement of faecal fat excretion as a test of fat malabsorption (and thus indirectly of pancreatic exocrine function), is now regarded as obsolete by most clinical biochemists and gastroenterologists.

Pancreatic enzymes that have been measured in faeces include chymotrypsin and elastase. Stool chymotrypsin measurements have suffered from a lack of standardization in the techniques used, making it difficult to compare results obtained from the various groups who have used the test. Measurement of faecal pancreatic elastase-1 is now recommended as the marker of choice for detecting pancreatic insufficiency. Elastase is an endopeptidase and sterol binding protein. As with chymotrypsin, elastase is not degraded during transit through the intestinal tract and is stable in faecal samples *in vitro*. Two commercially available enzyme linked immunosorbent assays (ELISAs) using antibodies specific for pancreatic elastase have been studied in patients with pancreatic insufficiency. Sensitivities of 60–100% for moderate to severe pancreatic disease, using a cut-off of 200 $\mu\text{g/g}$ wet weight, have been reported for faecal elastase. Discrimination between diarrhoea of pancreatic and non-pancreatic origins has been reported to be good, with better specificity than chymotrypsin. Faecal elastase measurements may also be useful in determining the amount of pancreatic enzyme

replacement therapy required in patients with cystic fibrosis or chronic pancreatic insufficiency (chymotrypsin cannot be used for this purpose). While biochemical tests attempt to give a functional diagnosis in chronic pancreatitis, imaging techniques provide further information that may disclose the cause. Plain abdominal radiography, endoscopic ultrasound, CT and magnetic resonance cholangiopancreatography (MRCP) are all useful. Endoscopic retrograde cholangiopancreatography (ERCP) with, or more commonly without, secretin stimulation is widely used, but carries a 10% risk of significant complications.

SMALL BOWEL BACTERIAL OVERGROWTH

With the discovery of microbes in the 19th century, theories were formulated linking the intestine with the development of systemic disease as a result of an unfavourable interaction between intestinal luminal microbes and the body. These ideas are equally evident today in the mass media, whereby producers of live yoghurts (containing viable bacteria) seek to equate ingestion of their products (probiotics) with health, vitality and happiness.

The normal intestinal microflora

Despite the perceived importance of intestinal bacteria, the reality is that we are still at the descriptive stages of assessing the normal intestinal flora and do not fully understand its function and impact on our well-being. Though the bulk of intestinal microbes inhabit the lower bowel, even the stomach is not usually sterile with bacterial population counts (predominantly Gram-positive aerobes) of up to 10^3 per gram of luminal content. The jejunum has a similarly Gram-positive flora, with a higher population of up to 10^4 per gram of luminal contents. The ileum contains a more varied flora that includes both aerobes and anaerobes, with populations in the region of 10^5 – 10^8 per gram. In the large intestine, Gram-negative anaerobes become the predominant species, overall bacterial populations rising significantly to 10^{10} – 10^{12} per gram of caecal content. Similar counts are present in the distal colon where the bacterial flora corresponds to that seen on faecal analyses.

The above data have been obtained using invasive techniques not commonly used in clinical practice. However, intestinal bacterial populations and species differ between individuals and relate in a complex way to age, diet, geographical and racial factors, antimicrobial treatment and intrinsic gut diseases.

Definition, causes and symptoms of small bowel bacterial overgrowth

The relatively small numbers of bacteria normally present in the small intestine are probably of little significance. However, problems can arise when the bacterial population of the small intestine is increased – bacterial overgrowth. An essential consideration of this definition is that bacterial overgrowth is based on quantitative and qualitative estimates of coliform bacteria in small bowel aspirates, something which is not possible in routine clinical practice. There are a number of diseases and conditions associated with small bowel bacterial overgrowth, a combination of factors predisposing to overgrowth in any given condition. Gastric hypochlorhydria of any cause may contribute to small bowel bacterial overgrowth. Other factors include: alterations in systemic immunology (e.g. isolated immunoglobulin A deficiency, hypogammaglobulinaemia, combined immune deficiency, infection with human immunodeficiency virus); impaired motility (e.g. advanced age, autonomic neuropathy in diabetes, fibrosis in scleroderma); pancreatic insufficiency; oral antibiotics; intrinsic small bowel disease (e.g. jejunal diverticulosis, coeliac disease, small bowel Crohn disease), and surgery (removal of the ileocaecal junction, surgically induced blind loops and, increasingly, bariatric surgery).

Small bowel bacterial overgrowth can be asymptomatic or be associated with non-specific symptoms, such as abdominal distension (bloating), eructation, flatulence, borborygmi and diarrhoea. The classic clinical presentation includes vitamin B₁₂ deficiency (intestinal bacteria compete with enterocytes for this vitamin), high blood folate concentrations (bacteria are a source of this vitamin) and fat malabsorption, and is rarely seen today. Rather, the condition is thought of when patients who are predisposed to overgrowth present with intestinal symptoms, for example diabetic patients with diarrhoea, weight loss and anaemia. An interesting suggestion is that small bowel bacterial overgrowth (intestinal dysbiosis) may be responsible for the symptoms in a proportion of patients diagnosed with IBS.

Diagnosis of small bowel bacterial overgrowth

Because of geographical and/or racial differences in small bowel bacterial flora, it is advisable that each laboratory independently establishes its own reference ranges for investigatory tests. The gold standard diagnostic procedure is microbiological examination (quantitative and qualitative bacteriological cultures) of small bowel contents, but this is impractical for routine clinical practice. More convenient, non-invasive biochemical procedures, based on oral administration of substances that are metabolized by bacteria to yield products that can be detected in breath,

have now been introduced, but their reliability remains controversial. The substances employed are either not absorbed from the small intestine (e.g. lactulose), therefore becoming available for metabolism by bacteria, or normally absorbed from the small bowel (e.g. glucose and D-xylose) and, therefore, only subject to metabolism by bacteria when these are present in the small bowel. Lactulose and similar test substances have the theoretical advantage of assessing the whole of the small bowel while the readily absorbed ones may be more selective for the proximal small bowel. The test results are also dependent upon the rate of intestinal transit. When transit is very rapid, the ingested load may be delivered to the caecum and give a false positive result. Furthermore, the tests vary in their sensitivity, presumably due to the influence of other ill-defined factors.

The ¹⁴C-glycocholate breath test was one of the first to be used for the non-invasive diagnosis of small bowel bacterial overgrowth, but it is obsolete because false negative results are common. Small bowel bacterial breath tests now in routine use include those involving test substances labelled with ¹³C (less commonly ¹⁴C), labelled CO₂ and breath hydrogen tests. Hydrogen breath tests depend on the fact that mammalian cells do not produce hydrogen. A breath hydrogen >20 ppm in expired gas after administration of the test substance (50g glucose or 10g lactulose) indicates bacterial overgrowth. Breath hydrogen is relatively simple to measure as a near-patient test. However, there are inherent limitations in the detection of breath hydrogen, for example with smoking and because of hydrogen production by oral flora. Carbohydrate may also be retained within the gut from an earlier meal. Abnormalities in gut transit and concomitant antibiotic therapy can also influence the result. Curiously, some individuals fail to produce breath hydrogen when challenged with a non-absorbed carbohydrate such as lactulose. There are a number of different protocols available for the performance of such breath tests and their value continues to be debated.

MALDIGESTION AND MALABSORPTION

It is conventional to distinguish between impaired uptake of nutrients due to maldigestion on the one hand and malabsorption on the other. The symptoms produced by the two conditions share many clinical features and there is a very broad spectrum of underlying causes. However, as the two processes are integrated and so inextricably linked, they are often lumped together as ‘malabsorption’. A very clear understanding of both the normal processes of digestion and absorption and the ways in which they can be disturbed is crucial to identifying the role that clinical biochemistry plays, both in making the clinical diagnosis and assessing prognosis.

Clinical features

Malabsorption can involve single or multiple nutrients. Accordingly, some patients may present with a single clinical feature; others may have several. Clinical features can relate to nutrient deficiency (or deficiencies), the

retention of nutrients in the gut (leading, e.g. to increased bacterial fermentation) or both.

Genetically determined intestinal lactase deficiency (see below) is a paradigm for maldigestion. The symptoms range from none, to those of abdominal bloating, excessive flatulence and watery diarrhoea (reminiscent of IBS), largely depending on the degree of the lactase deficiency, intake of lactose and intestinal transit time. Coeliac disease is an example of a condition causing malabsorption. Early recognition of the disease is now common but, in the past, the classic presenting features included: frequent loose stools, steatorrhoea, anaemia (iron or folic acid deficiency), wasting or impaired growth, osteomalacia (malabsorption of vitamin D and calcium), tetany (hypocalcaemia) and, occasionally, easy bruising (vitamin K deficiency). Exposure to gluten fractions present in wheat and several other cereals is an essential predisposing factor for coeliac disease. These fractions are thought to undergo an enzymatic reaction by transglutaminase (which catalyses the deamidation of specific residues of glutamine to glutamate). In patients with the disease (and mostly in those with the characteristic human leukocyte antigen (HLA) alleles DQ2 and DQ8), the peptide produced has an affinity for a groove in the encoded HLA molecules. This in turn evokes an abnormal intestinal mucosal T cell response, resulting in an inflammatory response that causes the characteristic changes of coeliac disease. The antibody response evoked can be useful in diagnosis. Anti-tissue transglutaminase antibodies are not a substitute for the definitive diagnostic test (the demonstration of total villous atrophy in a small bowel biopsy), but are useful as a screening procedure prior to endoscopy and in follow-up, as they usually become normal with successful treatment. Examples of ways in which malabsorption can present are shown in Table 12.2. Some of the basic laboratory investigations that should be carried out in patients in whom a diagnosis of malabsorption is suspected, are shown in Table 12.3. If all of these tests are normal, the diagnosis is unlikely; if they are abnormal, the next step is to look for a specific cause, guided by the history, clinical findings and laboratory results. Anaemia is

TABLE 12.2 Examples of the way in which malabsorption can present

Presentation	Cause
Passage of pale, bulky, offensive stools	Fat malabsorption or maldigestion
Abdominal distension, borborygmi, watery diarrhoea and excessive flatus	Malabsorption of carbohydrate
Vague malaise and tiredness, sometimes with weight loss (adults); growth retardation (children)	Generalized deficiency of nutrients
Anaemia	Deficiency of iron, folate or vitamin B ₁₂
Easy bruising or bleeding	Deficiency of vitamin K
Failure to thrive (infants)	Generalized deficiency of nutrients

TABLE 12.3 Some basic laboratory investigations in patients with suspected malabsorption

Investigation	To look for
Tests for evidence of malabsorption	
Full blood count	Anaemia, evidence of haematinic deficiency
Ferritin	Iron deficiency
Vitamin B ₁₂	B ₁₂ deficiency
Folate	Folate deficiency
Prothrombin time/ International normalized ratio	Vitamin K deficiency
Albumin	Malnutrition
Vitamin D	Vitamin D deficiency
Tests for causes of malabsorption	
C-reactive protein	Inflammatory conditions
Immunoglobulins	Hypogammaglobulinaemia
Antigliadin and anti-tissue transglutaminase antibodies	Coeliac disease
Faecal elastase	Pancreatic insufficiency

a particularly common feature of gastric and small bowel disease. The physiology and biochemistry of the vitamins and minerals involved in blood formation is discussed in detail in Chapter 27. Metabolic bone disease is discussed in detail in Chapter 31. In patients with malabsorption, much emphasis has hitherto been placed on malabsorption of calcium and vitamin D leading to osteomalacia. The advent of dual energy X-ray absorptiometry (DEXA) scans with reliable, sensitive and reproducible quantitation of bone mineral density has made it clear that there is a high prevalence of reduced bone mineral density in patients with coeliac disease and Crohn disease, even with successful treatment. Bone mineral density is frequently reduced in both disorders and osteoporosis may be seen in up to 25% of patients. Bone disease in these conditions predominantly affects the hips, with vertebrae and the forearm bones somewhat less severely affected. An important point is that reduced bone mineral density is often present in IBD at diagnosis and is independent of corticosteroid treatment (which is itself a cause of osteoporosis). Reduced bone mineral density in patients with IBD is associated with cytokine spillover from the gut into the circulation, leading to impairment of osteoblastic activity with no concomitant decrease in osteoclastic activity.

Carbohydrate absorption

Dietary carbohydrates

In most human diets, carbohydrates are the principal source of energy. The major form in the diet is starch, which accounts for about two-thirds of digestible dietary carbohydrate, the remaining third being made up of sucrose, lactose and their constituent monosaccharides. Starch is a polymer of glucose consisting of two forms: amylose and amylopectin. Amylose consists of long, unbranched chains of glucose molecules joined by α -1,4

links. Amylopectin consists of chains of glucose molecules with branching points for side chains every 12–25 glucose units. At each branching point, there is an α -1,6 link. Sucrose and lactose are disaccharides, sucrose consisting of a dimer of glucose and fructose, while lactose is a dimer of glucose and galactose.

Digestion of carbohydrates

Luminal events in carbohydrate digestion. The initial step in starch absorption is enzymatic digestion of α -1,4 links by salivary amylase to release maltose (dimer of two glucose molecules) and maltotriose (trimer of three glucose molecules). The salivary enzyme is rapidly inactivated by gastric acid, but hydrolysis is continued by pancreatic amylase, the end-products being maltose, maltotriose, short-branched oligosaccharides and α -limit dextrins, the latter two being residual branched segments resulting from incomplete digestion of amylopectin. No free glucose is produced.

Enterocyte events in carbohydrate digestion. The final step of carbohydrate digestion, the conversion of oligosaccharides to monosaccharides, is performed by

the disaccharidases of the small intestinal enterocytes. The main disaccharidases are maltase, sucrase-isomaltase and lactase (see Fig. 12.1). These enzymes are synthesized on the endoplasmic reticulum, transported to the Golgi apparatus and then to the brush border. They are distributed throughout the length of the small intestine, but sucrase and lactase are in highest concentrations in the jejunum. They are normally present in considerable excess, so that some reduction in activity does not normally result in symptoms.

The three main monosaccharides derived from the diet (glucose, galactose and fructose) are absorbed by saturable carrier-mediated transport systems in the enterocyte brush border. Secondary active transport of glucose and galactose occurs through a sodium co-transport system, driven by a sodium gradient dependent on Na^+, K^+ -ATPase on the basolateral surface of the cell. Fructose absorption is not active, but occurs through carrier-mediated (facilitated) diffusion.

Both the digestion and absorption of carbohydrates are highly efficient, but the process is incomplete, particularly for starch, with up to one-fifth of dietary starch failing to be absorbed. The efficiency of starch absorption varies, depending on the foodstuff from which it is

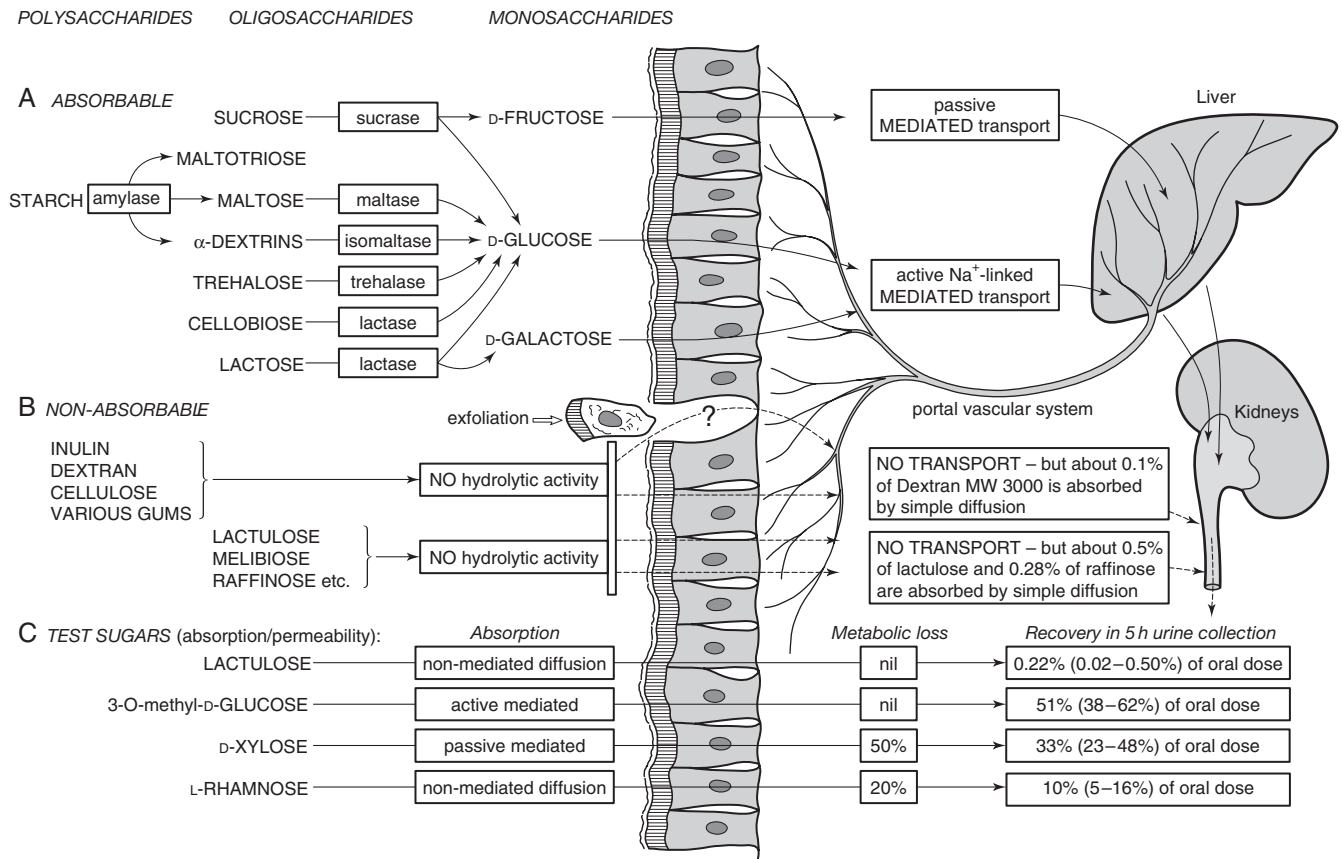


FIGURE 12.1 ■ Digestion and absorption of saccharides. (A) Ingested starch is broken down to the various disaccharides by salivary and pancreatic amylase. The disaccharides are cleaved by the various brush border disaccharidases to yield D-fructose, D-glucose and D-galactose. These in turn are transported across the brush border and enterocytes, and into the circulation by specific transport systems. (B) Non-hydrolysable polysaccharides, disaccharides and trisaccharides are thought to be excluded from permeation across the brush border. The small quantities that do permeate the intestine appear to pass by a paracellular route or through areas of cell extrusion (apoptosis). (C) Four commonly used sugars, used in a combined test of intestinal absorptive capacity and intestinal permeability, use different transport routes across the intestine, which allows assessment of the various intestinal functions. This increases the discriminatory value of the tests when assessing intestinal disease.

derived. For example, rice starch is highly absorbed, but between 10 and 20% of starch in wheat and certain beans enters the colon, where bacteria produce short chain fatty acids (which may be an important energy source for colonocytes), methane and hydrogen.

Clinical aspects of carbohydrate absorption

Although inherited disorders of sucrase-isomaltase and of transport proteins causing glucose-galactose malabsorption and fructose malabsorption are recognized, they are very uncommon. Adult lactase deficiency, however, is common in humans and may be regarded as a normal finding in many ethnic groups. Malabsorption of carbohydrates predisposes to osmotic diarrhoea with excessive flatus, abdominal distension and discomfort ('gripping' pains).

Lactase deficiency. Lactase activity is highest in the earliest months of life, but levels of the enzyme usually decline in all races after weaning. The prevalence of adult lactase deficiency is highly variable between races; so, for example, 5–15% of northern Europeans have demonstrable hypolactasia or alactasia, but >70% of Africans, Asians and especially Inuits have deficiency of the enzyme. However, the vast majority of affected individuals are asymptomatic, because overall completeness of disaccharide hydrolysis depends on the amount ingested, intestinal dilution and transit times, as well as the state of the enzyme (reduced or absent enzyme activity).

In practice, the diagnosis is often made by a therapeutic trial of a diet low in lactose, but metabolic tests are available (see below) and the diagnosis can be confirmed by assaying lactase activity in small bowel biopsy samples (though this is rarely required). Congenital lactase deficiency (i.e. lactase deficiency that is present at birth) has been described, but is extremely rare. The form of enzyme deficiency discussed above is usually known as primary or 'genetic' lactase deficiency. However, lactase deficiency may also complicate mucosal diseases of the small intestine and is then referred to as 'secondary'. This secondary lactase deficiency is a regular feature of coeliac disease, and is also seen in tropical sprue, IBD, radiation enteritis, chronic alcoholism with malnutrition and the enteropathy associated with acquired immune deficiency syndrome (AIDS). It may also accompany infections such as acute gastroenteritis and giardiasis, but usually resolves following resolution or successful treatment of the illness.

Investigation of carbohydrate absorption

The glucose tolerance test is of no value in the detection of malabsorption. The small intestinal reserve capacity for absorption of D-glucose is so great that absorption is usually well maintained, even when mucosal disease is severe. Furthermore, the rise in blood glucose concentration following oral administration is influenced by additional factors such as gastric emptying and endocrine status, which complicate interpretation.

Xylose absorption test. D-Xylose, a pentose sugar of plant origin, has been the basis of an absorption test

in routine clinical use since the 1930s. In the human, D-xylose is absorbed mainly from the jejunum, by a passive carrier-mediated system distinct from that utilized by D-glucose. Unlike D-glucose, there is no reserve capacity for the intestinal uptake of D-xylose, which made this sugar a potentially good indicator, responsive to minor changes in jejunal absorption. However, the concentration of xylose measured after an oral dose is affected by other factors including renal, hepatic and cardiac disease and the test is no longer widely available.

Lactose tolerance test. This test may be used in suspected lactase deficiency. In essence, serial blood samples for glucose estimation are taken after oral ingestion of 50 g of lactose. A rise in venous plasma glucose of <1.1 mmol/L is indicative of impaired lactose hydrolysis. The occurrence of typical symptoms following ingestion of lactose provides further evidence for the diagnosis.

Differential tests of intestinal disaccharide hydrolysis. Small amounts of non-hydrolysable disaccharides are absorbed by passive diffusion and appear in the urine: when hydrolysable disaccharides are not hydrolysed for some reason, the same thing happens. The intestinal uptake and urinary excretion of intact disaccharides in urine are, therefore, influenced by the extent of small intestinal disaccharidase activity: the greater the rate of hydrolysis, the less the excretion of intact disaccharide in urine. Excretion ratios of hydrolysable:non-hydrolysable disaccharides following ingestion of a mixture are, therefore inversely proportional to the efficacy of intestinal hydrolysis. Accurate, non-invasive assessment of intestinal lactase, sucrase or isomaltase can be made, using measurements of lactose/lactulose, sucrose/lactulose or palatinose/lactulose excretion ratios respectively, either separately or in combination. Ratios of lactose, sucrose or palatinose divided by lactulose (% molar dose excreted) may indicate isolated (genetically determined or primary) deficiency of lactase, sucrase (rare, a cause for severe diarrhoea in children) or palatinase (for isomaltase, not clinically relevant). However, if the absorption of two or more disaccharides is impaired, this suggests a secondary deficiency, for example mucosal diseases such as coeliac disease, infection or immune deficiency. Representative results using this test in patients with human immunodeficiency virus (HIV) infection are shown in [Figure 12.2](#).

This is undoubtedly the test of choice for the accurate and reliable assessment of whole bowel disaccharidase activity. It is, however, also more labour intensive and demanding on the patient than the other procedures mentioned above.

Protein absorption

The average Western diet contains some 80–100 g of protein per day, which provides not only the amino acids necessary for protein synthesis, but also some (10–15%) of the energy content of the diet. In addition, the gut digests and absorbs substantial amounts (perhaps 60 g) of endogenous protein derived from intestinal secretions, mucus and the shedding of mucosal cells.

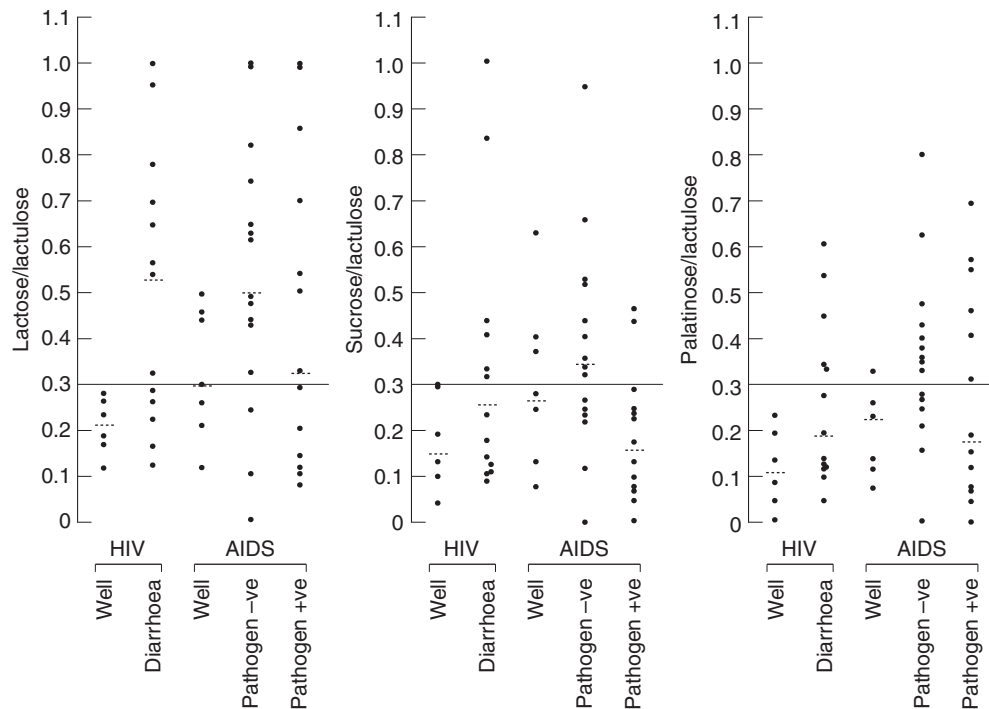


FIGURE 12.2 ■ Quantitative estimation of the rate of intestinal hydrolysis of ingested disaccharides in patients with HIV and AIDS. The test is carried out after an overnight fast and a preceding 12–18 h dietary exclusion of lactose- and sucrose-containing foods. Subjects ingest a solution containing lactulose 5 g, lactose 10 g, sucrose 10 g and palatinose 10 g (with or without L-rhamnose, which allows the integrity of intestinal permeability to be assessed at the same time), followed by a 10 h urinary collection with measurement of the disaccharides (% dose excreted) and calculation of the ratios. Well subjects have molar lactose/lactulose, sucrose/lactulose and palatinose/lactulose urine excretion ratios <0.3, indicative of normal intestinal lactase, sucrase and palatinase activities. Increased ratios in patients with the various stages of HIV are indicative of impaired hydrolysis of each of these sugars. The fact that the hydrolysis of all three sugars is impaired shows that this is secondary disaccharidase deficiency.

Digestion of proteins

Protein digestion takes place throughout the length of the small intestine and is efficient, with only 5% of the protein entering the gut lumen being lost in the stools each day.

Protein digestion begins in the stomach with the grinding and mixing action of gastric peristalsis and the food being exposed to gastric acid and pepsin. Pepsin is a protease that cleaves internal peptide bonds adjacent to hydrophobic amino acids. Pepsin is inactivated at a pH >4.0 and its action is, therefore, confined to the stomach. Digestion of proteins proceeds further through the action of the pancreatic proteases trypsin, chymotrypsin, elastase and carboxypeptidases A and B in the proximal jejunum. These are secreted as inactive proenzymes. Trypsin is released from trypsinogen by the brush border-derived enterokinase and activates the other proenzymes. Trypsin, chymotrypsin and elastase are endopeptidases, hydrolysing peptide bonds adjacent to certain specific amino acids, while carboxypeptidases A and B are exopeptidases, removing amino acids one at a time from the C-terminal ends of the peptides. The proteases reduce chyme to a mixture of free amino acids and oligopeptides (of between two and six amino acid residues). These are then further digested by brush border peptidases. There are at least eight mucosal peptidases that are synthesized and transported in a similar manner to the disaccharidases. However, unlike carbohydrates, the peptides are not completely digested to their constituent units, brush

border enzyme action leaving a mixture of amino acids, di- and tripeptides. There are specific transport systems that effectively absorb di- and tripeptides and, indeed, do so at a faster rate than for free amino acids; this discovery has had implications for the development of enteral feeding regimens. Peptide uptake is by sodium-linked active transport. Once inside the mucosa, peptidases within the enterocyte cleave small peptides to their constituent amino acids. A number of transport systems exist for the uptake of luminal amino acids and there are some rare inherited defects of amino acid absorption, including Hartnup disease (p. 171) and cystinuria (p. 170).

Clinical aspects of protein absorption

The most common cause of protein malnutrition worldwide is inadequate intake, often combined with poor absorption due to tropical enteropathy and frequent episodes of infective enteritis. The clinical syndrome is best illustrated by children from famine areas of the third world with their pot bellies (ascites because of hypoalbuminaemia). There are no specific syndromes of protein malabsorption, although protein deficiency can contribute to the wasting and nutritional disturbance of severe malabsorption syndrome. In several diseases, for example Crohn disease, ulcerative colitis and Whipple disease, excessive protein loss into the gut produces the syndrome of protein-losing enteropathy. This process is not one of malabsorption, although there may be associated malabsorption.

Investigation of protein absorption

There is no clinically useful non-invasive method of assessing protein digestion and absorption. Furthermore, there is little clinical need for such tests as mucosal diseases causing amino acid malabsorption are so rare. Protein-losing enteropathy can, however, be detected easily, most accurately by measuring the faecal excretion of intravenously administered albumin labelled with ^{51}Cr and somewhat less reliably by measurement of faecal α_1 -antitrypsin and α_1 -antichymotrypsin.

Fat absorption

Although the average fat content of the diet has been falling in the Western world over the last 20 years in response to medical advice, the average fat intake is still of the order of 90–100 g/day. There has also been a relative decline in the proportion of saturated fat in the diet with a concomitant increase in polyunsaturated fatty acids. The principal long chain polyunsaturated fatty acids are linoleic (C18:2) and linolenic (C18:3) acids. The other dominant fatty acids are oleic (C18:1) and palmitic (C16:0) acids. In food, these are mainly present as esters with glycerol (triacylglycerols or triglycerides). Additional dietary fat consists of phospholipid, cholesterol and fat-soluble vitamins. Because fats are insoluble in water, the mechanisms for their digestion and absorption are complex, but an understanding of the mechanism of fat digestion and absorption is crucial to an understanding of the wide variety of diseases that cause malabsorption of fat and are consequently associated with steatorrhoea.

Digestion of triacylglycerols

Luminal digestion. The first stage in the digestion of triacylglycerols is the process of peristaltic emulsification. Physical grinding of food, first by chewing and then in the stomach, produces an initial, relatively unstable emulsion that enters the duodenum. The emulsion is stabilized as the fat droplets are coated with phospholipid, provided exogenously in the diet and endogenously from bile. Further stabilization is achieved by a coating of bile salts and with small amounts of monoacylglycerols and fatty acid ions.

The main digestion of the emulsion occurs through the action of pancreatic lipase, but a minor initial component of digestion may be provided through the actions of lingual and gastric lipases, which have pH optima in the acid range. This component is important, however, in providing partial digestion of triglycerides to enhance and contribute to the stability of the emulsion. Pancreatic lipase acts in conjunction with a cofactor, pancreatic colipase, which is secreted in an inactive form (procolipase) and subsequently activated by tryptic digestion. Pancreatic lipase activity is greatest at near neutral pH and is therefore dependent on adequate bicarbonate secretion from the pancreas, under the influence of secretin. The combined action of lipase and colipase hydrolyses triacylglycerols to release fatty acids from the 1 and 3 positions, so that each triacylglycerol molecule is hydrolysed to a monoacylglycerol (with a fatty acid attached at the 2 position) and two

free fatty acids. Bile salts are amphipathic, that is, they contain both water-soluble and lipid-soluble components. This property allows them to aggregate in micelles such that the hydrophobic components line up adjacent to one another on the inside of the micelle with the hydrophilic aspects facing outwards into the aqueous phase. The water-insoluble fatty acids, monoglycerides and cholesterol are, therefore, held 'inside' the micelle to form a highly stable particulate emulsion. While triacylglycerol absorption is fairly efficient in the absence of bile salts, by contrast the absorption of cholesterol and fat-soluble vitamins is severely compromised unless the bile salt concentration in the duodenum is above a critical micellar concentration.

Absorption of triacylglycerols. The transport of lipids across the surface of enterocytes is by passive diffusion, that is, the lipids partition across lipid components in the brush border membrane. The microclimate at the enterocyte surface is slightly more acidic than in the jejunal lumen, which encourages the liberation of fatty acids from micelles. There is some suggestion, however, that a component of fat absorption for certain fatty acids, particularly linoleic and oleic acids, may occur by facilitated diffusion. Triacylglycerols are resynthesized from monoacylglycerols and fatty acids in enterocytes and incorporated into chylomicrons. The carrier protein apolipoprotein B48 is essential for chylomicron synthesis, and absence of this protein, as in congenital abetalipoproteinaemia, produces severe fat malabsorption. While longer chain triacylglycerols (above C10) pass into the lymphatics, medium chain triacylglycerols (C6–10) are at least partly absorbed intact, probably by passive diffusion, to reach the liver by the portal circulation.

Digestion and absorption of other fats

Phospholipids are hydrolysed to fatty acids and lysophosphatidylcholine by pancreatic phospholipase A₂, while cholesteryl esters are hydrolysed by a pancreatic esterase. The fatty acids form micelles with bile acids, and thus both lipids are ultimately transported to enterocytes in a similar way to the digestion products of triacylglycerols. Phospholipids are resynthesized in the endoplasmic reticulum of mucosal cells and cholesterol is esterified by acyl-CoA cholesterol acyltransferase before being incorporated into chylomicrons. The efficiency of fat absorption is considerable so that, in health, the process is virtually complete. Most fat absorption occurs within the proximal jejunum.

Clinical aspects of fat malabsorption

Fat absorption is a complex, multistage process. Therefore, disease at several levels in the alimentary system can result in fat malabsorption, which, when severe, is characterized by the presence of steatorrhoea. Interference with delivery of bile salts into the duodenum, pancreatic exocrine deficiency, failure to achieve near-neutral duodenal pH either from hypersecretion of gastric acid or hyposecretion of pancreatic bicarbonate, and mucosal disease due to villous atrophy, lymphangiectasia, abetalipoproteinaemia etc, can all result in steatorrhoea. Tests of fat absorption have been used in the past as screening methods to detect malabsorption. In modern clinical practice, it

is increasingly evident that confirmation of steatorrhoea (rather than quantification) by microscopy of faeces for fat after suitable staining is sufficient to initiate appropriate investigation of the relevant organs that may play a role in fat digestion and absorption.

Investigation of fat absorption

Faecal fat excretion. Quantitative measurement of faecal fat excretion is the most sensitive method of detecting fat malabsorption. However, this investigation is unpleasant both for patients and laboratory staff, and few laboratories now include it in their repertoires. There have been a number of efforts to find alternative measures of fat absorption. The ^{14}C -triolein (or ^{13}C) breath test is probably the best of these, but, nevertheless, is not widely used.

$^{13/14}\text{C}$ -triolein breath test. Following an overnight fast and collection of a basal sample of expired air, the patient is given ^{13}C - or ^{14}C -triolein in a 60 g fat meal. Hourly breath samples are then collected for 7 h for measurement of labelled carbon expiration. If triolein is effectively absorbed, large amounts of labelled CO_2 will appear in the expired breath, fat malabsorption being indicated by smaller amounts being exhaled. However, when steatorrhoea is macroscopically obvious, confirmation by measurement is not required, and investigation should be directed towards diagnosing the cause.

INTESTINAL PERMEABILITY

An alternative approach to the measurement of absorptive function in the assessment of small bowel disease is to assess barrier function – that is the ability of the intestine to exclude compounds from being absorbed. Increased intestinal permeability is not characteristic of any single disease, but is seen in a variety of gastrointestinal pathologies.

Previously, tests using the differential permeability of combinations of sugars were used to assess intestinal permeability, but these are now largely obsolete, because of capsule enteroscopy. For those interested in these tests, details are provided in Further reading.

FAECAL TESTS OF INTESTINAL INFLAMMATION

The intestinal tract is a most hospitable environment for bacteria. The intestine has developed a very effective and complicated way of accommodating them and protecting the host, but it is clear that the defence system is somewhat less than 100% efficient. If there is a breach in mucosal integrity of any sort, the microbes gain an advantage and elicit a significant inflammatory reaction. Hence, almost all colonic and most small bowel disorders are associated with inflammation.

The first reliable quantitative non-invasive test for assessing intestinal inflammation was the ^{111}In white cell technique, which involved abdominal scintigraphy (for localization of disease) and a four-day faecal collection (for quantitative estimation of the inflammatory activity).

The test is uniquely sensitive for the detection of inflammation, but is not disease-specific. Disadvantages are its cost, the requirement for special labelling facilities, a high dose of radiation and the inconvenience of a complete four-day faecal collection. Simpler methods were needed and subsequently developed. The qualities of a marker for assessing intestinal inflammation are that it should be stable, resistant to degradation by intestinal bacteria, easily extractable from faeces, easily assayed and, preferably, specific for a particular type of inflammatory cell. Two such markers have undergone rigorous testing, namely calprotectin and lactoferrin with other proteins, including M2PK and S100A12, entering the field more recently. Calprotectin acts as a representative for this group of proteins and its use will, therefore, be described in detail.

Calprotectin

Calprotectin accounts for about 60% of total soluble proteins in the cytosolic fraction of neutrophil granulocytes. It is a member of the S100 family of calcium and zinc binding proteins (a heterodimer of S100A8/A9) and is found in neutrophils, monocytes and some squamous epithelial cells. It is released by activation of leukocytes as a consequence of inflammatory disease. Calprotectin has antimicrobial properties and, in addition, it can inhibit the proliferation of both normal and malignant cells, probably by sequestration of zinc, which is a critical element for many enzymes. In particular, the metalloproteinases MMP-2, MMP-3, MMP-7, MMP-8, MMP-9 and MMP-13 are inhibited *in vitro* by calprotectin at biologically relevant concentrations, suggesting that it may have important regulatory functions.

In the presence of calcium, calprotectin is remarkably resistant to proteolytic degradation and it is this fact that underlies its stability in faeces. Commercially available kits can be obtained for the extraction of calprotectin from faeces and quantitation by ELISA and require no more than a 200 mg sample. Reference ranges should be established by individual laboratories, as there are almost certainly racial and geographical variations. In the UK, the upper limit of the reference range for faecal calprotectin is 50–60 $\mu\text{g/g}$. African Caribbeans resident in the UK may have values up to 200 $\mu\text{g/g}$ without disease. In general, when used as a screening procedure for intestinal disease, increased values between 50 and 200 $\mu\text{g/g}$ rarely disclose significant disease. In these cases, we recommend repeating the test before further investigation. Calprotectin concentrations between 200 and 500 $\mu\text{g/g}$ are commonly seen in patients with diverticulitis, colon malignancies and polyps, and those on NSAIDs, whilst those with values >500 $\mu\text{g/g}$ almost invariably have IBD or intestinal infections.

Calprotectin in disease

Inflammatory bowel disease

Studies from different groups have shown that faecal calprotectin concentrations are almost invariably increased above healthy controls in patients with active ulcerative colitis and Crohn disease. The clinical implication is that faecal calprotectin clearly has potential as a screening test to distinguish patients with IBS, who will have normal

concentrations, from those with IBD, whose concentrations are characteristically increased by at least ten-fold in active disease.

Patients with clinically inactive IBD may have a normal faecal calprotectin (depending on treatment), but often have an increased concentration. The considerable overlap between the values in clinically active and inactive disease is such that it is inadvisable to make predictions of a patient's clinical status based solely on faecal calprotectin concentrations.

Faecal calprotectin is particularly useful in patients with inactive ulcerative colitis and Crohn disease, as values five times above the upper limit of normal have 90% sensitivity and 83% specificity in predicting clinical relapse within six months, allowing the opportunity for early treatment.

Colorectal cancer

Colorectal cancer (adenocarcinoma) is the second most common cause of death from malignancy in the Western world. Survival rates are closely related to the stage of cancer at the time of diagnosis, and the most promising approach to reducing mortality rates is the early detection of precancerous or cancerous lesions. The most widely accepted non-invasive method for detecting colorectal cancer is faecal occult blood (FOB) testing. Screening in asymptomatic populations has, at best, reduced mortality rates by 15–33%; a national bowel cancer screening programme using faecal occult blood testing was initiated in the UK in 2006.

Various studies have consistently demonstrated that faecal calprotectin concentrations are increased in patients with colorectal cancer. Sensitivity is of the order of 80–90%, whereas that for FOB is only 40–58%. Furthermore, faecal calprotectin is increased in 50% of patients with colorectal polyps (some of which may be precancerous). However, the test is not recommended for cancer screening purposes because of its lack of specificity. Numerous faecal molecular marker probes, usually polymerase chain reaction (PCR) based, are currently being tested for the detection of colorectal cancer. To-date, they lack sensitivity, but are very specific.

Irritable bowel syndrome

Given that faecal calprotectin has a sensitivity of almost 100% in detecting patients with active IBD, this raises the possibility of using it as a screening test to distinguish between these patients attending outpatient clinics and the more common patients with IBS who do not warrant invasive investigation. The main value of the test in patients with IBS is that a normal test result practically rules out chronic inflammatory bowel disease and thereby reduces the need for invasive investigation. Of patients with IBS, 15–20% have raised faecal calprotectin, typically to 2–3 times normal.

Neuroendocrine tumours of the gastrointestinal tract and pancreas (NETs)

Neuroendocrine tumours of the gastrointestinal tract are relatively rare, but easily diagnosed when they produce clinically distinct and specific syndromes. However, many

of the tumours do not secrete substances that lead to the specific syndromes, and are only detected because of the physical effects of local or metastatic lesions. A proportion of these tumours have a genetic predisposition due to a loss of tumour suppressor genes, namely multiple endocrine neoplasia (MEN) type 1 and 2 (see Chapter 41).

As a group, the neuroendocrine tumours are characterized histologically by sheets of small round cells that are derived from intestinal neuroendocrine cells in the submucosa, with very infrequent mitosis. It is, therefore, not usually possible to distinguish between a benign and malignant tumour histologically, unless there is evidence of metastasis. This contrasts very sharply with the common gastrointestinal cancers (oesophageal, gastric and colonic) that originate from the epithelial cells (squamous cell cancer and adenocarcinomas), where histopathological diagnosis is usually straightforward. From the clinical biochemistry viewpoint, it is most logical to discuss the different tumours according to the clinical syndromes that they cause.

Intestinal carcinoid tumours and the carcinoid syndrome

The carcinoid syndrome. The cardinal presenting symptoms of the carcinoid syndrome are persistent diarrhoea (40–80%), flushing (30–70%) and nondescript abdominal pain (10%). The diarrhoea is watery and the need to defecate often coincides with the flushing periods. The flushing is intermittent, of sudden onset, brief (lasting for <15 min), associated with a feeling of warmth and can be precipitated by alcohol ingestion, exercise or by certain foods, especially cheese. The flushing is mainly on the neck and face and is otherwise not distinctively different from menopausal flushing or drug-induced flushing (chlorpropamide and calcium channel blockers). Moreover, there are other well-recognized, albeit less common, manifestations of the carcinoid syndrome. Cardiac valvular involvement (10% of patients) present as right sided heart failure. A further 10% of patients develop asthma-like symptoms, and those with metabolically active tumours that utilize much of the ingested tryptophan may have pellagra-like skin lesions due to nicotinic acid (niacin) deficiency.

The carcinoid syndrome is a manifestation of a carcinoid tumour, a tumour of argentaffin cells that can occur anywhere in tissue derived from the embryonic gut, but most frequently is found in the ileum. Its pathophysiology, diagnosis and management are discussed in detail in Chapter 41. Measurement of urine 5-HIAA and/or serum chromogranin A is used in the diagnosis and monitoring of carcinoid syndrome patients.

Pancreatic endocrine tumours

Pancreatic endocrine tumours have been variously called 'islet cell tumours' or APUDomas (amine precursor uptake and decarboxylation) and are of similar origin to carcinoids. They share certain features, consisting of homogeneous sheets of small cells with very few mitoses. Malignant potential may be indicated by local invasion or metastasis, and each tumour is usually associated with

symptoms attributable to a single hormone/peptide, although the tumour may release a number of hormones. Unlike carcinoids, pancreatic endocrine tumours are mostly functional, that is, they are associated with symptoms that are attributable to the hormone that they release. Their prominence in textbooks is disproportionate to their clinical importance, and the vast majority of patients are managed and treated within specialist units. As a group of tumours, they are exceptionally rare with an annual incidence of 1–2/million population. Insulinomas and gastrinomas are the most common, with vasoactive intestinal peptide (VIPomas), glucagonomas and somatostatinomas being ten times rarer, while others are confined to single case reports. Pancreatic endocrine tumours are not infrequently discovered because of their association with MEN-1 and 2 (see Chapter 41) or the phacomatoses (tuberous sclerosis, von Hippel–Lindau and von Recklinghausen disease).

Insulinomas. Tumours of the β cells of the pancreas secrete insulin without the normal physiological feedback action of circulating blood glucose. Fifty per cent of insulinomas are <1 cm in diameter at diagnosis, most are solitary (except when associated with MEN) and 90% are benign. Insulinomas typically present with symptoms of hypoglycaemia, especially after fasting or exercise. The most common symptoms are loss of consciousness (sometimes presenting as coma), dizziness, diplopia, blurred vision, confusion, abnormal behaviour, excessive sweating, palpitations and weight gain (because of the increased food intake that relieves some of the symptoms). The diagnosis is based on the presence of ‘Whipple’s triad’, namely symptoms of hypoglycaemia, documented hypoglycaemia and relief of symptoms when glucose is administered to restore normal blood glucose concentrations.

The first step in the diagnosis of insulinoma is the documentation of fasting hypoglycaemia, which can best be differentiated from postprandial hypoglycaemia from the clinical history. A low plasma glucose concentration in the presence of inappropriately high insulin and C-peptide concentrations is strongly suggestive of an insulinoma. Insulin and C-peptide are formed in equimolar amounts from the enzymatic cleavage of proinsulin. A high insulin concentration in the presence of a low C-peptide concentration is indicative of the administration of exogenous insulin. Hypoglycaemia associated with misuse of sulfonylureas may need to be excluded from the differential diagnosis by measurement of serum sulfonylurea concentrations. Other causes of hypoglycaemia, including alcohol misuse, Addison disease and liver failure, should also be excluded. Hypoglycaemia can be life-threatening and only surgery can offer curative treatment. Techniques such as CT and MRI may localize tumours larger than 1 cm, but at least 50% of tumours are smaller than this. Angiography has a high detection rate for smaller insulinomas, as these are highly vascular, and endoscopic ultrasound is increasingly used for their detection. In the most difficult cases, it may be necessary to sample venous blood selectively from the pancreas in order to detect a differential increase in C-peptide concentrations following arterial injection of calcium.

Glucagonoma. The clinical hallmark of a glucagonoma is the symptom complex of migratory necrolytic erythema, weight loss and anaemia. Glucose concentrations may be high due to the antagonist action of glucagon to insulin. The skin lesions are the most distinct component of glucagonoma, and the difficulty recognizing them may account for the fact that they have often been a feature of the disease for an average of six years prior to diagnosis.

Unlike insulinomas, glucagonomas are usually large (>5 cm) at the time of diagnosis and >50% are malignant. Localization by conventional imaging techniques is, therefore, not a problem. The diagnosis can be confirmed by finding raised plasma glucagon concentrations: these are usually elevated five-fold or more. Minor elevations of glucagon can be found in chronic renal failure, acute pancreatitis and severe trauma, especially with burns and sepsis, but these conditions are easily identified clinically. Surgery should always be considered despite the malignant nature of many of the tumours. The medical treatment is largely directed at control of diabetes; octreotide is sometimes useful.

VIPoma. VIPomas are rare tumours, characterized by copious watery diarrhoea with dehydration and hypokalaemia. The severity of the diarrhoea is comparable to that seen in cholera with excretions in excess of 2 L/24 h. The diagnostic feature of the diarrhoea is that it continues despite prolonged fasting. Plasma VIP concentrations are elevated 4–5-fold; the same imaging techniques are used to localize the tumour as in glucagonomas. VIPomas have the same propensity as glucagonomas for malignancy. While surgery is usually considered, their medical management remains important with rehydration and correction of electrolyte disturbances in severe cases, and use of antidiarrhoeals (e.g. loperamide, codeine phosphate and somatostatin analogues).

Somatostatinoma. Excessive secretion of somatostatin (the somatostatin syndrome) is characterized by diabetes (somatostatin inhibits release of insulin); diarrhoea (because of inhibition of exocrine pancreatic function); weight loss and gallstones (a result of impaired gallbladder contraction). However, as pancreatic surgery for tumours becomes more common, it is evident that only 10% of somatostatinomas are functional, that is, associated with the somatostatin syndrome. Somatostatinomas tend to be large (2–10 cm at diagnosis), and the majority are malignant with evidence of metastasis being present at the time of diagnosis. A significant number are detected during cholecystectomy or following pancreatic tumour work-up. Occasionally, symptomatic patients with the somatostatin syndrome are diagnosed on the basis of increased plasma somatostatin concentrations. Treatment options include surgery and chemotherapy.

THE ACUTE ABDOMEN

Introduction

The term acute abdomen is used to indicate the rapid onset of severe symptoms that may indicate potentially life-threatening intra-abdominal pathology requiring urgent

surgical intervention. Its clinical importance is not in doubt. Many of the diseases that present in this way have a high mortality if not treated surgically; conversely, others are associated with high mortality and morbidity if surgery is performed. The vast majority of patients with an acute abdomen present initially to an accident and emergency department (at least in the UK) and are then transferred to the care of surgeons.

The clinical diagnosis is based principally on the nature of the pain. Common causes of the acute abdomen are shown in [Box 12.1](#). Additionally, there are diseases such as measles causing ileitis and perforation that have a higher incidence in the developing world. Non-abdominal pathologies may simulate an acute abdomen.

History and examination often yield a likely diagnosis, but investigations are helpful in providing confirmation, e.g. testing for pregnancy in women with lower abdominal pain, measurement of amylase when acute pancreatitis is suspected. In addition, a full blood count may reveal an increase in neutrophil count suggestive of infection or anaemia secondary to blood loss, and it is important to have knowledge of renal function and electrolytes prior to undertaking surgery.

BOX 12.1 Causes of an acute abdomen

Common causes

- Non-specific abdominal pain
- Acute appendicitis
- Acute cholecystitis, ascending cholangitis
- Small and large bowel obstruction
- Intussusception, volvulus and strangulated hernias
- Renal/ureteric colic
- Perforated peptic ulcer
- Acute pancreatitis
- Acute diverticulitis
- Gynaecological conditions (salpingitis, ectopic pregnancy)
- Ruptured aortic aneurysm
- Bleeding (haematoma) associated with anticoagulation

Less common causes

- Gastroenteritis including infection with *Salmonella*, *Shigella*, *Yersinia*, measles virus etc.
- Crohn disease
- Mesenteric lymphadenitis
- Pyelonephritis
- Meckel diverticulitis

Conditions that may simulate an acute abdomen

- Acute intermittent porphyria
- Myocardial infarction
- Myocarditis
- Pulmonary embolism
- Pneumothorax
- Pneumonia
- Oesophagitis (reflux, infectious)
- Oesophageal rupture
- Sick-cell crisis
- Acute leukaemia
- Herpes zoster
- Familial Mediterranean fever

There follows a brief account of the disorders where clinical biochemistry may play an important role in confirming the diagnosis of diseases that present as an acute abdomen, namely acute pancreatitis, ectopic pregnancy and porphyria. It should, however, be appreciated that other diagnostic modalities can also provide important information, for example, ultrasound or CT examination in suspected acute pancreatitis.

Acute pancreatitis

Acute pancreatitis is an acute inflammation of the pancreas. Some cases may be self-limiting, but others may be severe and associated with organ failure including pulmonary insufficiency and acute kidney injury and with local complications such as pancreatic necrosis, abscess or pseudocyst formation.

[Table 12.4](#) shows the factors that are recognized to be important in its aetiology, of which gallstones and alcohol account for >80% of cases. The precise pathogenesis is often uncertain, but the most widely accepted hypotheses suggest that insults to the pancreas cause activation of zymogens (inactive pro-enzymes) within the gland and consequent autolytic destruction and production of inflammatory cytokines and chemokines. How the initial insult brings about these changes is disputed, although ischaemia may promote both.

Almost all patients with acute pancreatitis have abdominal pain. This is severe epigastric pain, usually of sudden onset, often with radiation to the back. In more severe cases, there is nausea and vomiting, fever, hypotension, shock and multiorgan failure that may lead to death. Biochemical features of acute pancreatitis include: uraemia (compromised renal function); hypoalbuminaemia (extravasation); hypocalcaemia (partly related to hypoalbuminaemia and partly because fatty acids released in and around the inflamed pancreas by pancreatic lipase form insoluble calcium salts); hyperglycaemia (increased sympathetic activity and destruction of the islets of Langerhans); metabolic acidosis; hypoxaemia, and raised plasma activities of liver enzymes. None of these is invariably present and none is diagnostic for pancreatitis. In acute necrotizing pancreatitis, methaemalbuminaemia may be detectable. The main use of these tests is to provide information about the prognosis, and they may also influence management. Biochemical tests that are used for their diagnostic potential in acute pancreatitis include amylase and lipase.

Amylase

In acute pancreatitis, amylase activity rises within 2–12 h of the onset of symptoms, usually returning to normal within 3–5 days. Serum amylase above five times the upper reference limit is very suggestive of acute pancreatitis. Smaller increases may have other explanations (see [Table 12.1](#)).

In exceptionally rare cases, pancreatitis severe enough to cause death may occur without a rise in serum amylase. A persistently high amylase following an attack of acute pancreatitis suggests the formation of a pancreatic

TABLE 12.4 Causes of acute pancreatitis

Factors	Comment
Gallstones	The most frequent, and best established, aetiological factor. Migrating small gallstones may block the pancreatic duct (some people believe that reflux of bile into the pancreas is also important) and thereby initiate the damage
Alcohol misuse	The acute effects of ethanol on the pancreas are complex. Alcohol alters pancreatic blood flow and, coupled with the generation of free radicals from ethanol metabolism, this may cause free radical damage. Alcoholic stimulation of pancreatic secretion coupled with spasm of the sphincter of Oddi could lead to obstructive injury. Ethanol may also sensitize the pancreas to the effects of other agents that lead to zymogen activation and the generation of harmful cytokines. The overall contribution of these factors to the development of alcoholic pancreatitis is unclear
Drugs	Thiazide diuretics, oral contraceptives, corticosteroids, 5-aminosalicylic acids (used in the treatment of inflammatory bowel disease), azathioprine, 6-mercaptopurine, asparaginase, didanosine, valproate, statins and other drugs have all been implicated as aetiological agents, but only rarely. The mechanisms are unknown
Hypercalcaemia	Often quoted, but the majority of patients with hypercalcaemia do not develop pancreatitis
Hypertriglyceridaemia	Severe hypertriglyceridaemia is important in its own right, but excessive alcohol ingestion is a confounding factor in many cases
Trauma	Blunt trauma to the abdomen may occasionally cause acute pancreatitis; the trauma may also be iatrogenic, e.g. following endoscopic retrograde cholangiopancreatography (ERCP)
Infectious causes	Coxsackievirus, <i>Ascaris lumbricoides</i> , <i>Candida</i> , <i>Salmonella</i> , HIV
Rare causes	Pancreatic tumours, hereditary pancreatitis, autoimmune pancreatitis, scorpion toxins, idiopathic

pseudocyst, a cyst containing pancreatic enzymes, which needs to be differentiated from an abscess, which is infective in nature. Since amylase is excreted in urine, the diagnostic efficacy of urinary amylase measurement in acute pancreatitis has also been investigated. Diagnostic performance can be improved by the simultaneous measurement of amylase and creatinine in paired serum and urine samples. The normal ratio of amylase clearance to creatinine clearance is 2.2–4.2%, rising to 6.3–13.3% in acute pancreatitis. This approach is less convenient than simply measuring serum amylase. In general, the main use of urinary amylase is in establishing a diagnosis of macroamylasaemia. This is a condition where serum amylase is consistently raised (characteristically 2–3-fold) owing to the formation of complexes of amylase from non-pancreatic and pancreatic origin with immunoglobulins, usually IgG. When this is the case, a raised serum activity is not accompanied by a correspondingly increased urinary amylase. The isoenzyme raised in the plasma in illness depends on the source of the amylase, for example in pancreatitis, the P enzyme is raised, whereas in ruptured tubal pregnancy, the S isoenzyme is increased, and in renal failure, both S and P are raised. An increase in both isoenzyme fractions of amylase with a normal lipase is also suggestive of macroamylasaemia.

Lipase

The rise in plasma lipase activity approximately parallels that of amylase. Measurement of both enzymes may improve diagnostic accuracy, at least in part because the fall in lipase is slower, so that it remains raised for longer following an acute attack. Lipase activity increases in any condition where hyperamylasaemia is due to pancreatic pathology, but not when the amylase is of non-pancreatic origin. In particular, serum lipase activity is normal in macroamylasaemia. While lipase measurements overcome some of the non-specificity of amylase estimation in acute pancreatitis, lipase may still be raised

in non-pancreatic disease. This, together with the fact that analytical methods for lipase are more complex and expensive than those for amylase means that lipase measurements are not widely used in the UK. Nevertheless, national guidelines recommend the measurement of lipase for diagnosis.

Choice of test for pancreatitis

Of the approaches discussed above, the measurement of serum amylase activity is the most frequently used biochemical test for pancreatitis, mainly because it is easy to perform. In areas where there is a large population of African Caribbeans (who have a higher reference range), the measurement of lipase may occasionally be useful to establish whether a raised amylase is of pancreatic origin. A number of other markers of acute pancreatitis have been investigated (e.g. elastase and phospholipase A₂), but as yet these have not become established in routine use.

Ectopic pregnancy

A fertilized ovum occasionally may implant outside the uterus, where it gives rise to an ectopic pregnancy. The most common site is in one of the Fallopian tubes. Although this is not a common cause of an acute abdomen, even in young women, it is important because if a tubal pregnancy ruptures, the resulting massive intra-abdominal haemorrhage is life-threatening.

Serum progesterone is lower in absolute concentration and human chorionic gonadotrophin (hCG), which normally doubles in concentration every two days after implantation, tends to increase somewhat less rapidly in ectopic than in normal pregnancy. However, while such measures can be useful, they are not applicable in an acute situation. Rapid, sensitive urinary hCG assays are now available as near patient tests and will usually indicate whether or not the patient is pregnant (although occasionally hCG is undetectable in both maternal urine and

serum in ectopic pregnancy). Quantitation of serum hCG is valuable, since, in a normal pregnancy, a gestational sac should be detectable in the uterus on transabdominal ultrasonography once the hCG is >1000 IU/L. The clinical diagnosis of ectopic pregnancy can be difficult, and the results of biochemical tests have to be considered in light of the clinical assessment and the results of other techniques, such as transabdominal and transvaginal ultrasonography.

Acute porphyria

Acute porphyria is an exceptionally rare cause of an acute abdomen. The diagnosis may be suggested by the history (e.g. the association of the abdominal pain with the ingestion of a drug known to provoke acute attacks in susceptible subjects), or occasionally by someone noticing that the patient's urine is dark red, or that it darkens on standing.

The porphyrias are discussed in detail in Chapter 28, but in a patient with acute abdominal pain, a negative screening test for urinary porphobilinogen (PBG) effectively excludes porphyria as the cause of the pain. (δ -Aminolaevulinic acid dehydratase deficiency remains a theoretical possibility, but only isolated cases of this condition have been reported.) If the test for urinary PBG is positive, then one of the acute porphyrias is likely.

Further reading

- Ayling RM. New faecal tests in gastroenterology. *Ann Clin Biochem* 2012;49:44–54.
- Bjarnason I, Macpherson A, Hollander D. Intestinal permeability: an overview. *Gastroenterology* 1995;108:1566–81.
A review of the clinical aspects of intestinal permeability testing.
- Bures J, Cyrany J, Kohoustova D et al. Small intestinal bacterial overgrowth syndrome. *World J Gastroenterol* 2010;16:2978–90.
Review of aetiology, pathogenesis, diagnosis and treatment of small intestinal bacterial overgrowth syndrome.
- Feldman M, Friedman LS, Brandt LJ, editors. Sleisenger and Fordtran's gastrointestinal and liver disease. 9th ed. Philadelphia: WB Saunders; 2010.
This is the single most important and comprehensive textbook of gastroenterology available and contains a most extensive, accurate and up-to-date account of biochemistry and the clinical biochemistry tests available, covering practically all gastrointestinal diseases in detail.
- Goddard AF, James MW, McIntyre AS et al. Guidelines for the management of iron deficiency anaemia. *Gut* 2011;60:1309–16.
British Society of Gastroenterology guidelines for investigation and management of IDA.
- Lieb II JG, Draganov PV. Pancreatic function testing: here to stay for the 21st century. *World J Gastroenterol* 2008;14:3149–58.
Review of tests for chronic pancreatitis with particular emphasis on the combined biochemical and endoscopic tests.
- Ramage JK, Ahmed A, Ardill J et al. Guidelines for the management of gastroenteropancreatic neuroendocrine (including carcinoid) tumours (NETs). *Gut* 2012;61:6–32.
UK and Ireland Neuroendocrine Tumour Society guidelines for NETs including the use of biochemical tests and imaging for diagnosis.
- Thomas PD, Forbes A, Green J et al. Guidelines for the investigation of chronic diarrhoea, 2nd ed. *Gut* 2003;52(Suppl. V):v1–15.
The British Society of Gastroenterology guidelines on tests for malabsorption.

Assessment of hepatic function and investigation of jaundice

Roy A. Sherwood • Adrian Bomford

CHAPTER OUTLINE

INTRODUCTION 231

ANATOMY OF THE LIVER 232

- The hepatic circulation 232
- Macroscopic structure 232
- Microscopic structure 232
- Ultrastructure 233
- Bile, bile ducts and biliary drainage 234

HEPATIC REGENERATION 234

PHYSIOLOGICAL FUNCTIONS 234

- Carbohydrate metabolism 234
- Lipid metabolism 234
- Protein metabolism 235
- Biotransformation and excretion 235
- Bile secretion 236

LIVER FUNCTION TESTS 236

- Bilirubin and bile pigment metabolism 237
- Plasma enzyme activities 239
- Plasma proteins 241
- Bile acids 243
- Quantitative evaluation of liver function 243

- Other tests of liver function 244

USES OF LIVER FUNCTION TESTS 245

- Differential diagnosis of jaundice 245
- The inherited hyperbilirubinaemias 245
- Monitoring response to therapy 246
- Neonatal jaundice 247

ABNORMAL LIVER FUNCTION TESTS IN ASYMPTOMATIC PATIENTS 247

- Bilirubin 247
- Alkaline phosphatase 247
- Aminotransferases 247
- γ -Glutamyltransferase 248

NORMAL LIVER FUNCTION TESTS IN THE PRESENCE OF OVERT LIVER DISEASE 248

ROLE OF LIVER FUNCTION TESTS IN ASSESSING PROGNOSIS 248

- Chronic liver disease 248
- Acute liver failure 249

CONCLUSION 249

INTRODUCTION

It is widely accepted that the term 'liver function tests', as currently used in clinical practice, is a misnomer. Certainly this is true as compared with pulmonary or renal function tests. These describe the measurement of distinct physiological and biochemical organ functions that have meaning in the absence of any organ pathology. In marked contrast, most of the parameters that comprise the standard liver 'function' tests, such as the plasma activities of certain aminotransferases, which have major and well-defined roles within the cell, have no *functional* significance at all in plasma, where they are simply markers of hepatocyte disruption. Their measurement is only of any significance when applied to liver pathology. To this extent, understanding of the conventional liver function tests relies on a broad grasp of the principles of liver disease, and the present chapter should, therefore, be read in conjunction with Chapter 14.

This is not to say that the function of the liver is not well understood. But, while it is conventional to list the functions that the liver can perform, this detracts from gaining a broad conceptual picture of what the liver 'does'. Essentially, the liver is a regulatory barrier between the systemic circulation and the organism's environment experienced via the gut. The job of the acinus, the functional unit of the liver, is to regulate the concentrations of solutes entering the systemic circulation via the terminal hepatic venules or being excreted in the bile. It is the principal organ of metabolic homeostasis, that is, maintenance of blood composition within physiologically acceptable limits by the conversion, synthesis and release of components required by other organs and by removal of toxic substances that may be injurious to tissues.

This chapter reviews briefly the anatomy, physiology and biochemistry of the normal liver as a basis for understanding the tests currently applied in clinical practice and those that may be developed in the future.

ANATOMY OF THE LIVER

The macroscopic and microscopic anatomy of the liver is difficult to understand, partly because of its inherently complicated three-dimensional structure and partly because of the recent trend to replace simple (but misleading) morphological descriptions with more accurate, but less obvious, functional descriptions.

The hepatic circulation

The liver has a dual blood supply. Arterial blood, direct from the aorta, is supplied via the hepatic artery from the coeliac axis. The second source is the portal vein, which is formed by the joining of the superior mesenteric and splenic veins and which collects blood from the gut. After passage through the sinusoids (see below), blood drains from the liver via the hepatic veins at the posterior aspect of the liver into the inferior vena cava and thence to the right side of the heart. The portal venous system delivers about 80% of the blood and 20% of the oxygen supplied to the liver. A major cause of abnormal function in chronic liver disease is disturbance of blood flow through the liver, consequent on the fibrosis that follows chronic liver cell damage, but if the portal vein is blocked, relatively normal liver function can be maintained for many years. Arterial occlusion occurring suddenly, for example as a result of trauma, leads to acute liver failure, but more gradual occlusion by a tumour mass is compatible with normal liver function for quite long periods.

Macroscopic structure

The liver is a wedge-shaped organ located in the right upper quadrant of the abdomen. Its mass varies with that of the individual, being in the order of 22 g/kg body weight. In a typical 70 kg subject, the liver weighs about 1.5 kg. It has a large right lobe, a smaller left lobe anteriorly and two further small lobes, the quadrate and the caudate lobes. These lobes relate to the venous drainage, not to the portal distribution (see below). Thus, the left hepatic vein drains the left hepatic lobe and the right and middle hepatic veins drain the right hepatic lobe. In terms of the portal structures, there are two *functional* lobes defined by the right and left portal veins. The division is marked by a line joining the inferior vena cava and the gallbladder bed (Fig. 13.1).

Microscopic structure

Hepatocytes, the hepatic parenchymal cells, comprise about 80% of the total cell mass of the liver. As viewed under the microscope, the functional unit of the liver appears to be the acinus, often termed the lobule, and it is in terms of this structure that pathological changes are described. Since liver biopsy has, under many circumstances, become the *de facto* 'gold standard' of liver disease diagnosis, it is important to understand this terminology. It is also important to understand that this unit does not comprise a homogeneous collection of hepatocytes and

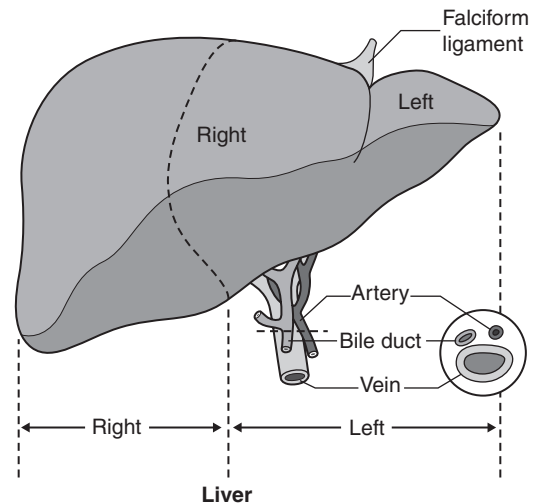


FIGURE 13.1 ■ The anterior surface of the liver. The labelling in the upper part of the figure represents the conventional description into a right and left lobe separated by the falciform ligament. The lower part of the labelling refers to the right and left lobes defined by distribution of the portal structures. The right and left portal structures (portal vein, hepatic artery and bile ducts) enter the functional right and left lobes, respectively.

that there is marked functional heterogeneity across the acinus, with different zones having different physiological and biochemical functions (see below).

The acinus

The acinar/lobular concept of the structure of the liver is based on a central efferent terminal hepatic venule surrounded by radially oriented plates of hepatocytes and sinusoidal channels. Three to five portal tracts (or 'triads'), containing branches of the portal vein, hepatic artery and bile duct, are situated around the periphery of each acinus (Fig. 13.2A).

The acini represent microcirculatory units comprising groups of liver cells, each fed by a single terminal portal venule and hepatic arteriole, the blood from which passes via the sinusoids into a terminal hepatic vein (Fig. 13.2B). Flow is unidirectional, from portal tract to hepatic vein. The 20 or so hepatocytes that separate the portal tracts from the terminal hepatic vein in an acinus have been arbitrarily separated into three zones, through which the portal blood passes sequentially. Zone 1 is conventionally used to describe hepatocytes nearest the afferent arteriole in the portal tract (the periportal area), and those surrounding the terminal hepatic vein, in the 'centrilobular' or 'perivenular' area, are described as zone 3. Between these two zones is an ill-defined area (zone 2), which is intermediate in terms of the composition of the perfusing blood. It must be emphasized that there are no anatomical boundaries between these three zones but, clearly, as blood perfuses each zone sequentially, its composition is altered and this leads to the functional heterogeneity of hepatocytes across the acini in response to the changes in their microenvironment. Zone 1 hepatocytes receive blood rich in oxygen and have a high level of metabolic activity. It is not surprising, therefore, that oxidative functions of the liver tend

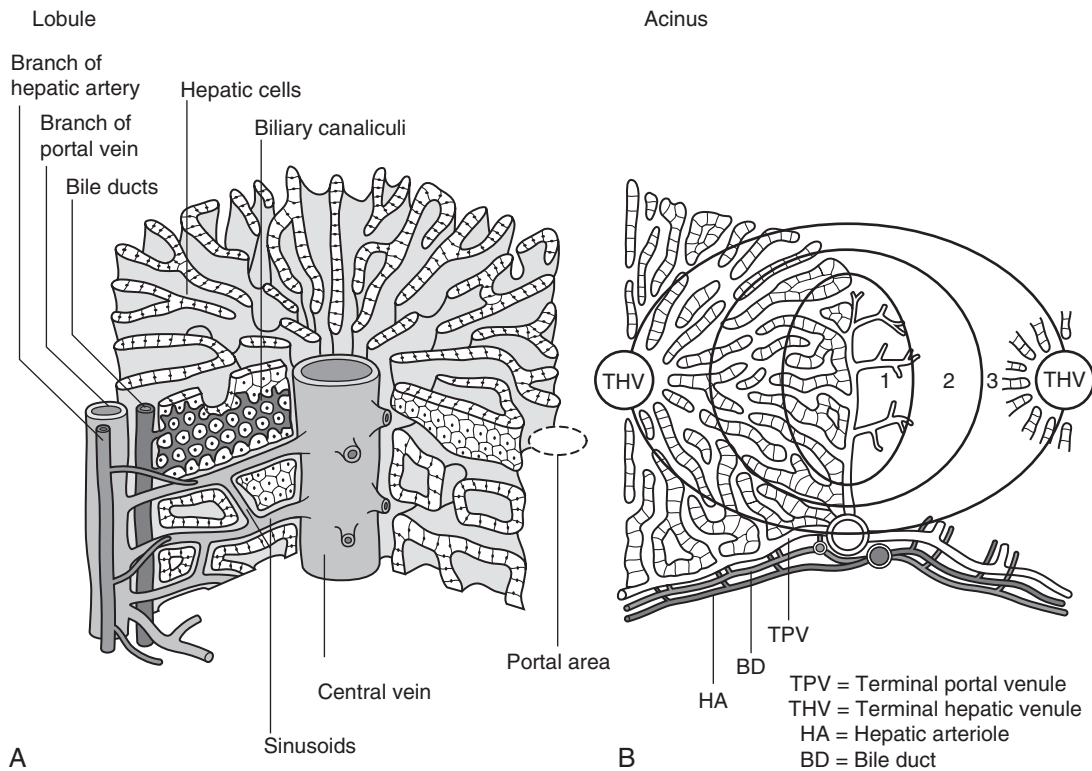


FIGURE 13.2 ■ (A) The conventional hepatic acinus (lobule) based on the central vein with surrounding portal tracts. (B) The hepatic acinus and its three functional zones. The axis is formed by the terminal portal venule, hepatic arteriole and bile duct. Blood flows from the periphery of the acinus to the terminal hepatic venule.

to be performed mainly by hepatocytes in zone 1, with metabolic processes that can operate at lower oxygen tensions occurring mainly in zone 3.

A detailed discussion of the functional heterogeneity of the three zones is not undertaken here, but we mention some examples that could be important if damage to the liver is not uniform. Thus, oxidative metabolism (the respiratory chain, citric acid cycle and fatty acid oxidation), gluconeogenesis, urea synthesis and the production and excretion of bile all occur mainly in zone 1, while glycolysis, glutamine synthesis from ammonia, and xenobiotic metabolism occur predominantly in zone 3. It is not difficult to appreciate that different pathological insults may differentially damage the various zones. Perhaps blood tests may eventually be developed to investigate the integrity of the different zones and thereby reduce the need for histological examination in the diagnosis of different liver disorders.

Ultrastructure

Hepatocytes are arranged in single-cell sheets or 'plates', supported by a fine mesh of a collagenous material (reticulin) and separated from overlying fenestrated endothelial lining cells by the space of Disse (Fig. 13.3). The hepatocytes are exposed to blood flowing through the sinusoids on each side of the plates. Their sinusoidal surfaces have a microvillar structure that greatly increases the surface area of the cell membrane, thereby facilitating efficient exchange of solutes between the blood and the cells. Bile produced by the hepatocytes is excreted via specific transporters located in the membrane of the biliary canaliculi (see below). The latter are formed by

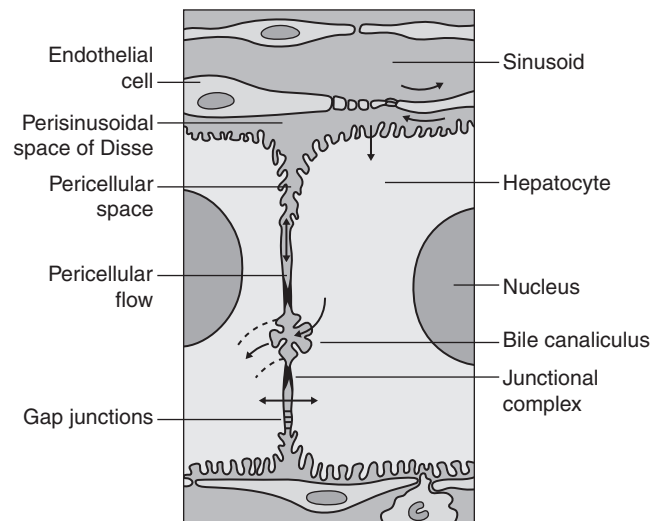


FIGURE 13.3 ■ Ultrastructure of the liver. Solute transport is shown across the sinusoid, which does not have a conventional basement membrane, and across the space of Disse. Here they may be taken up across the hepatocyte membrane and subsequently across the canalicular membrane, or enter the canaliculus through the 'paracellular pathway' via the intercellular junctions.

invaginations of the smooth basolateral membranes between adjacent hepatocytes, and also have a microvillar structure. Direct communication between the blood and bile is prevented by tight junctions in the basolateral membranes on either side of the biliary canaliculi. In addition to the endothelial cells that filter fluid exchanged between the sinusoidal lumen and

the space of Disse, the sinusoids also contain Kupffer cells (a major part of the reticuloendothelial system), hepatic stellate cells (vitamin A-storing, fat-storing cells, also known as Ito cells, that produce several connective tissue components) and liver-resident natural killer cells with antitumour activity. These non-parenchymal cells are an area of intensive research; current evidence suggests that they play a major role in synthesis of growth factors responsible for control of liver regeneration and development of fibrosis in liver disorders (see below).

Bile, bile ducts and biliary drainage

The biliary canaliculi formed between adjacent hepatocytes are continuous with canaliculi between other hepatocytes within the plane of the liver cell plates, and eventually drain into ductules lined by specific biliary epithelial cells and ultimately into the major bile ducts, thence to the common bile duct and the gut. Interruption of the flow of bile, the exocrine secretion of the liver, is responsible for many of the signs of hepatobiliary disease. Bile pigment and bile acid metabolism are described below, and the pathological anatomy of the liver and biliary tract is described in Chapter 14.

HEPATIC REGENERATION

The fact that the liver has a great capacity for regeneration is the rationale for much of the practice of clinical hepatology, particularly for undertaking prolonged periods of liver-intensive care during acute liver failure and for the feasibility of extensive resection, particularly for the management of primary and, less commonly, secondary malignant liver tumours.

Hepatic regeneration has been studied mainly in animals, most extensively the rodent model, after partial hepatic resection. Twenty-four hours after liver resection, there is extensive division of the remaining cells accompanied by a surge in DNA synthesis; this is followed 24 h later by increased replication of the non-parenchymal cells. The initiation of these events is by growth factors that are now being identified. Epidermal growth factor (EGF), transforming growth factor α (TGF α) and hepatocyte growth factor (HGF) are all involved in switching on regeneration, and transforming growth factor β (TGF β) is involved in switching it off. Normally quiescent differentiated hepatocytes replicate rapidly after resection, while intrahepatic precursor cells, termed oval cells, proliferate and generate lineage only in situations where hepatocyte proliferation is blocked or delayed. Bone marrow cells can generate oval cells, but such transdifferentiation is now known to be inefficient and occurs only rarely.

PHYSIOLOGICAL FUNCTIONS

Because the various functions of the liver may fail at different times, creating differing clinical pictures, it is necessary to consider each function separately. The physiology and

biochemistry of the liver subsume most of intermediary metabolism and, as such, are clearly beyond the scope of this chapter. Other than giving a brief outline, emphasis is placed on those functions whose measurement or disturbance is important in liver tests or pathology, respectively.

Carbohydrate metabolism

As has been known for many years, removal of the liver in an animal leads to death from hypoglycaemia. A major metabolic function of the liver is to store sugar and reduce the variations in blood glucose consequent upon the human habit of eating intermittently. Thus, during a meal, the liver stores glucose as glycogen and then releases it (glycogenolysis) slowly when food is not being eaten. This is particularly important for those organs such as the brain and red blood cells that have an obligatory requirement for glucose. Specific glucose transporter molecules located in the sinusoidal membrane that mediate facilitated diffusion are involved in both processes. Between meals, as the supply of glycogen decreases (only about 75 g can be stored), the liver starts to make glucose (gluconeogenesis) from other sources, particularly lactate, but also pyruvate, glycerol and alanine, although only from alanine is there a significant *net* production of glucose. During more prolonged periods of starvation, the total body requirement for glucose falls and energy demand is increasingly met by production of ketone bodies, derived mainly from fatty acids via acetyl-CoA. Apart from being stored, glucose is used by the liver as an energy substrate via glycolysis and the citric acid cycle, or for the synthesis of fatty acids and triglycerides.

Insulin is secreted in response to the rise in blood glucose concentration after a meal and promotes an increase in peripheral glucose uptake and a decrease in gluconeogenesis. In acute liver failure, the liver may not be able to maintain an adequate concentration of blood glucose and hypoglycaemia may become a life-threatening complication; in *chronic* liver disease, hyperglycaemia is more common, most likely because of a failure of the liver to store glycogen and failure of peripheral tissues to take up glucose adequately. The liver also metabolizes other dietary sugars including fructose and galactose, converting them to glucose phosphates (see Chapter 14).

Lipid metabolism

After a meal, dietary triglycerides are hydrolysed to free fatty acids and monoglycerides by pancreatic lipases and dissolved in an aqueous medium, facilitated by the detergent action of bile salts excreted by the liver into the gut. The liver meets its own metabolic energy requirements, and those of the body as a whole, by mitochondrial β -oxidation of short chain fatty acids. The resultant acetyl-CoA either enters the citric acid cycle or reacts with another molecule of acetyl-CoA to form ketone bodies. Although the role of the liver is central to the oxidation of fatty acids, most tissues contain the enzymes required to undertake complete oxidation. The liver also synthesizes fatty acids, triglycerides, cholesterol, phospholipids and lipoproteins. Disturbances of fatty acid metabolism, including decreased oxidation (as in excessive alcohol

consumption), increased hepatic fatty acid synthesis and decreased breakdown of triglycerides to fatty acids and glycerol, may all be involved in the development of fatty liver ('steatosis'), an increasingly prevalent problem found in obesity and as an additional or incidental histological finding in many liver disorders.

Protein metabolism

Hepatic protein metabolism is central to the assessment of liver function, and its disturbance underlies many of the clinical complications that occur in severe liver cell dysfunction.

Synthesis

Other than the immunoglobulins, most circulating proteins are synthesized wholly or largely by the liver, and the concentrations of several are used as a measure of hepatic synthetic function. Apart from albumin, transcobalamin II and C-reactive protein, all are glycoproteins. Glycosylation (often with terminal sialic acid residues on the carbohydrate moieties) has several functions. In some instances (e.g. fibronectin), it serves to make the protein resistant to proteolysis; in others it affects function. In yet others (e.g. caeruloplasmin), it affects the half-life of the protein in the blood, because hepatocytes possess receptors that can bind galactose and some other carbohydrate residues exposed after removal of sialic acid and remove the desialylated glycoproteins from the circulation. Disturbance in glycosylation produces some specific defects in protein structure that may be useful clinically in the diagnosis of alcoholic (see p. 242) and malignant liver disease (see Chapter 14).

Metabolism of amino acids and disposal of urea

A 70 kg man on a normal diet needs to excrete between 10 and 20 g of nitrogen per 24 h. This derives, in the form of ammonia, from amino acids that are surplus to requirements (and cannot be stored) and those that are not reutilized after normal turnover. The ammonia is converted into urea in the liver and excreted by the kidneys. The liver processes dietary amino acids arriving via the portal vein and from breakdown of muscle proteins, both for its own requirements and for export to peripheral tissues. Aromatic amino acids (AAA: phenylalanine, tyrosine and tryptophan) are metabolized by the liver, but hepatic extraction of branched chain amino acids (BCAA: leucine, isoleucine and valine) is small and these are taken up largely by muscle. The ratio of BCAA/AAA is decreased in acute liver failure and this alteration forms the basis of one theory of the pathogenesis of hepatic encephalopathy, namely that it is due to the toxic effects of increased concentrations of ammonia on the brain.

The major pathways of ammonia production and clearance are shown in Figure 13.4. Amino acids first undergo transamination to glutamate, followed by oxidative deamination with the formation of ammonia. The resultant ammonia is fed into the Krebs–Henseleit (urea) cycle and excreted as urea or stored transiently as glutamine

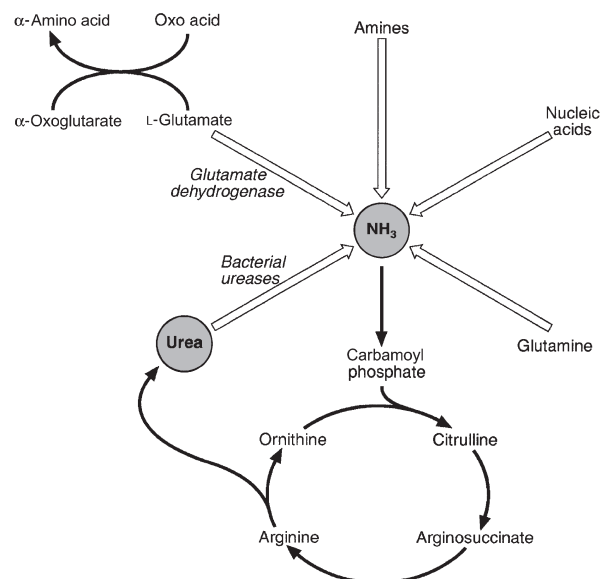


FIGURE 13.4 ■ Production and disposal of ammonia.

(through the action of glutamine synthase). Additional ammonia is produced by the action of bacteria on dietary protein and urea present in gut contents. Plasma ammonia concentration is measured in many laboratories and a raised concentration is taken as evidence that an encephalopathic state is due to hepatic pathology. Measurement of the various enzymes involved in the urea cycle is increasingly undertaken for diagnosis of inherited abnormalities of urea synthesis.

Biotransformation and excretion

The lungs and kidneys are effective in excreting volatile and water-soluble substances, respectively, but many compounds, both exogenous (such as drugs) and endogenous, including end products of metabolism, are lipid soluble and non-volatile. Many of these compounds are toxic, and a vital function of the liver is to render such substances more water soluble so that they can be excreted in urine or bile.

Two phases of biotransformation of metabolites by the liver are recognized. In phase I, a suitable polar group is made available, which is conjugated in phase II. Phase I reactions occur in the smooth endoplasmic reticulum and are mediated mainly by the mixed function oxidase system (cytochrome P450 isoenzymes) that utilize atmospheric oxygen, typically generating hydroxylated or carboxylated compounds. Phase II reactions involve their subsequent conjugation by the action of glucuronyl transferases with glucuronic acid, acetyl or methyl radicals or, in the case of bile acids, with glycine, taurine or sulphate (see below). There is considerable functional heterogeneity of the glucuronyl transferases, of which there are several isoenzymes that have varied substrate specificities, particularly for exogenous compounds. An alternative, non-oxidative pathway for biotransformation of both endogenous and exogenous substances is by conjugation with reduced glutathione by glutathione S-transferases.

Bile secretion

The constituents of bile, the exocrine secretion of the liver, are conjugated bile salts, cholesterol, phospholipids (mainly lecithin), bilirubin mono- and diglucuronides, electrolytes and small amounts of protein. The liver is the major site of cholesterol biosynthesis and the sole site of conversion of cholesterol into bile acids, which are the major organic anions excreted by the liver. The primary bile acids, cholic and chenodeoxycholic acid, are conjugated with either glycine or taurine to form bile salts, which enhances their solubility at the pH of body fluids (Fig. 13.5). This facilitates their main function of solubilizing both biliary cholesterol and the products of dietary fat digestion resulting from lipid hydrolysis. Secondary bile acids, deoxycholic and lithocholic acids, are derived from primary bile acids by the action of intestinal bacterial 7α -dehydroxylase, either as bile salts or deconjugated bile acids. Most of the bile acids reaching the gut are reabsorbed in the terminal ileum and return via the portal vein to the liver (enterohepatic circulation) where, as free acids, they are again conjugated and excreted. A third metabolite, ursodeoxycholate (a stereoisomer of chenodeoxycholic acid), is found in trace amounts and has been classified as a tertiary bile acid. This enterohepatic circulation is regulated by the activities of distinct bile salt transport proteins, including the canalicular bile salt export pump, the ileal sodium-dependent bile salt transporter and the hepatic sinusoidal sodium-taurocholate co-transporting polypeptide. Several other bile salt transporters and organic anion-transporting polypeptides have been characterized (see below). During bile secretory

failure (cholestasis), bile salt transport proteins undergo adaptive responses that serve to protect the liver from retention of toxic bile salts and facilitate non-hepatic routes of bile salt excretion. The measurement of serum bile acids has been extensively investigated as a test of liver function (see p. 243).

LIVER FUNCTION TESTS

Liver function tests have four potential applications.

1. *As an aid to establishing whether an individual has liver disease.* In effect, they are being used to answer the question, 'Is there any evidence of liver damage or dysfunction?'
2. *As an aid to making a specific diagnosis.* While functional tests are clearly distinct from diagnostic tests, it is still reasonable to assume that certain patterns of dysfunction may be characteristic of particular diseases.
3. *To establish the severity of liver dysfunction or damage once a specific diagnosis has been established.* This is important from a prognostic point of view, although the standard biochemical 'liver function tests' do not always reflect accurately the severity of tissue damage.
4. *To monitor the progression of the disease and any response to therapeutic intervention.*

Within this framework, two classes of tests will be considered. The first are the standard liver function tests – a group of tests often applied irrespectively of the suspected

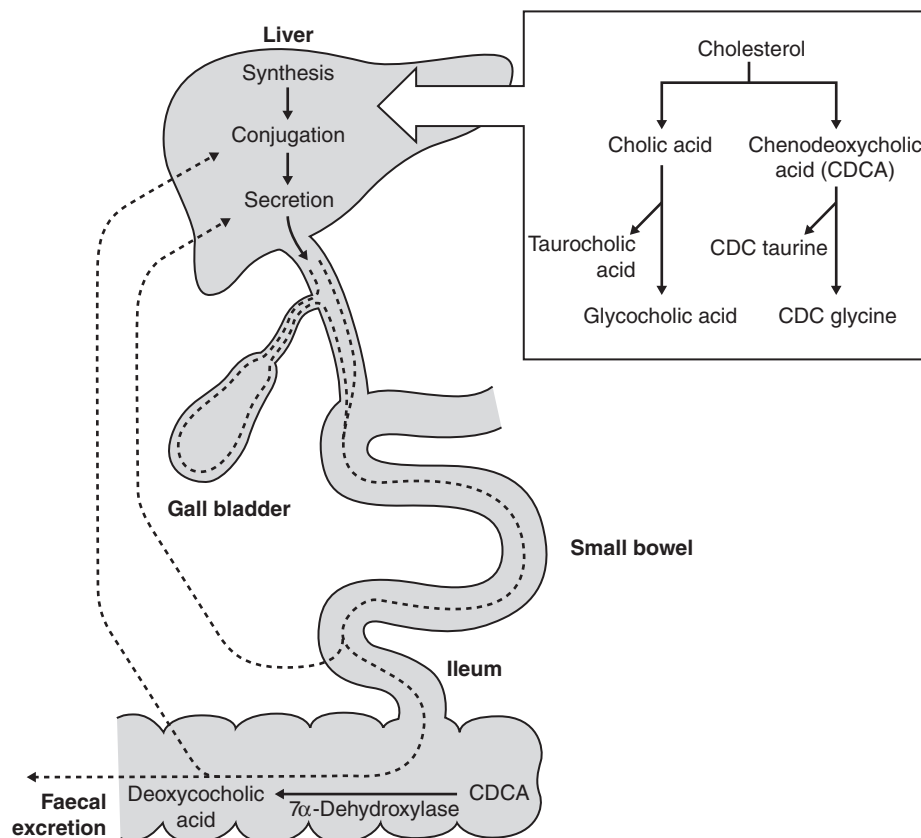


FIGURE 13.5 ■ Synthesis and enterohepatic circulation of bile acids.

diagnosis and to which all the caveats about lack of true functional assessment apply. The second group includes those biochemical assays used for assessment of liver disease in specific situations, for example α_1 -antitrypsin in suspected deficiency of this protein and α -fetoprotein in suspected primary hepatic cancer (hepatocellular cancer). The latter are described briefly here, for the sake of completeness, and in more detail in the next chapter in the context of their diagnostic application. The standard liver function tests are usually considered to include the plasma total bilirubin concentration, the activities in plasma of certain enzymes (particularly alanine and/or aspartate aminotransferases, alkaline phosphatase and γ -glutamyltransferase), and the total plasma protein, albumin and globulin concentrations. The prothrombin time (discussed later) also provides useful information about the synthetic capacity of the liver in the short term, as it is a functional measure of factor VII (among others) concentration and this clotting factor has a short half-life in plasma (see below). It has been estimated that this group of tests will correctly allocate patients to a liver disease/non-liver disease category in about 75% of cases.

Bilirubin and bile pigment metabolism

Although bilirubin has been characterized as a non-toxic metabolic product of a relatively minor metabolic pathway, marked elevation in its plasma concentration leads to the alarming sign of jaundice and usually suggests the presence of underlying liver or biliary tract disease that may range from trivial to life-threatening, especially in neonates (see below). Accurate interpretation of the laboratory tests associated with bile pigment metabolism in the jaundiced individual requires a clear understanding of the physiology and biochemistry of the bile pigments. Most bilirubin is derived from the breakdown of haem, itself derived from senescent red blood cells (Fig. 13.6). A much smaller proportion comes from other haemoproteins such as catalase, myoglobin and the cytochromes. An even smaller fraction comes from 'ineffective erythropoiesis', although this may represent a significant source of bilirubin in haematological conditions such as thalassaemia and pernicious anaemia. The initial and rate-limiting step is the oxidation of haem to biliverdin by haem oxygenase; this is followed by reduction (catalysed by biliverdin reductase) to bilirubin, with the production of an equimolar amount of carbon monoxide and ferric (Fe^{3+}) iron. These reactions take place in the macrophages of the reticuloendothelial system, predominantly in the liver, spleen and bone marrow. The resultant 'unconjugated' bilirubin is tightly bound to albumin in a 1:1 molar ratio, but additional binding sites of lower affinity are recruited in hyperbilirubinaemic states. This binding limits extrahepatic uptake of the potentially toxic unconjugated bilirubin and facilitates transport to the liver. Other molecules, such as thyroxine and certain drugs, can compete for albumin binding sites and thereby displace bilirubin, although the clinical relevance of this displacement is limited, except possibly in neonates.

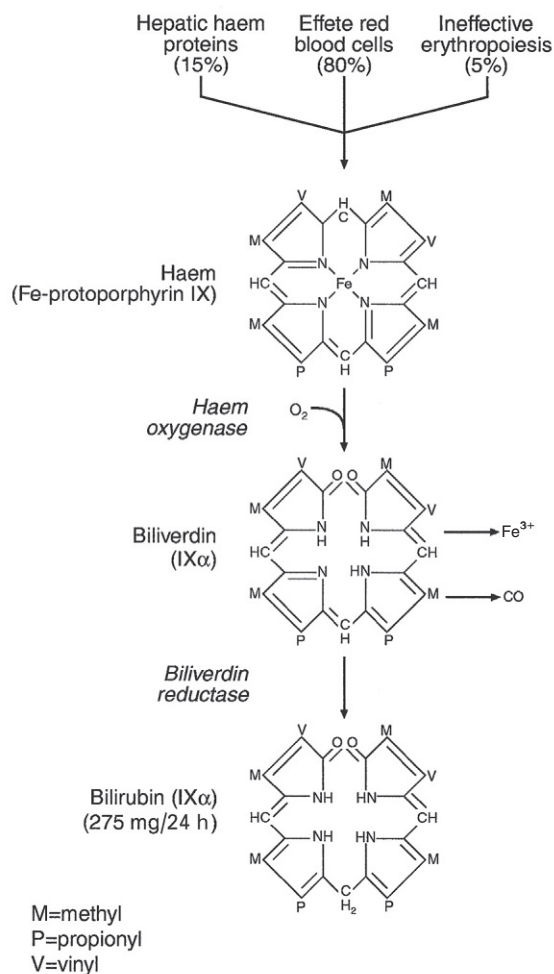


FIGURE 13.6 ■ Formation of bilirubin.

Bilirubin, tightly bound to albumin, is actively transported across the hepatocyte sinusoidal membrane and binds to ligandin (glutathione transferase B). It is then conjugated with glucuronic acid by the action of uridine diphosphate (UDP)-glucuronosyltransferase (glucuronyl transferase) to form mono- and diglucuronides and thereby rendered water soluble (Fig. 13.7). Glucuronidated bilirubin is secreted via an active transport mechanism into the biliary canaliculi and thence reaches the gut. The components of this transport pathway have been identified as products of genes encoding multidrug resistance (MDR) proteins (MRPs) and MDR-associated proteins. The protein MDR3 is known to transport phosphatidylcholine to chaperone bile salts, while MRP2 pumps glucuronidated compounds, that is, conjugated bilirubin as well as organic anions. Secretion is probably rate limiting in the overall transport of bilirubin from plasma to bile. In the gut, some bilirubin is deconjugated by bacterial glucuronidases and (being fat soluble again) is reabsorbed, but most is oxidized to urobilinogen; this is further metabolized to other pigments, particularly stercobilin, and excreted (Fig. 13.8). A small amount of urobilinogen is reabsorbed and undergoes an enterohepatic circulation. Conjugated bilirubin also circulates bound to albumin, but with low affinity, such that the unbound fraction can

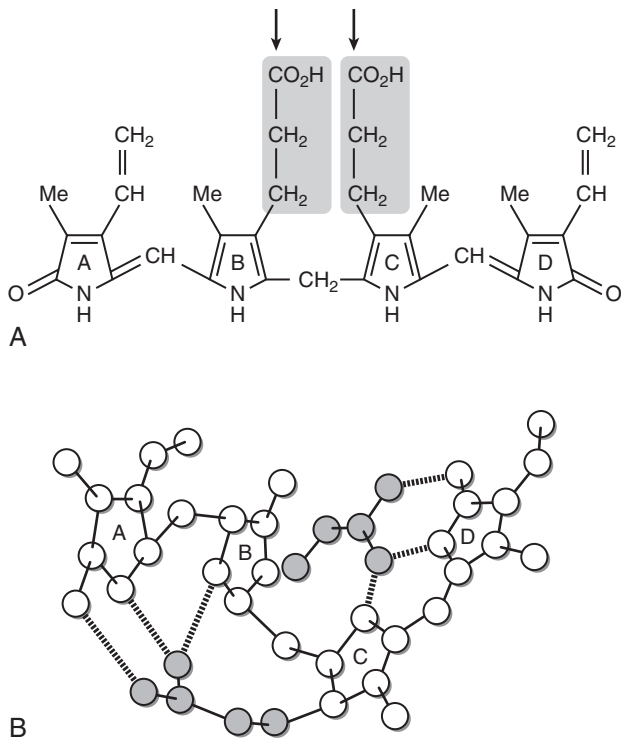


FIGURE 13.7 ■ (A) The unfolded structure of bilirubin showing the site of glucuronic acid conjugation (arrows), which breaks the hydrogen bonding and results in the molecule becoming soluble. (B) The folded structure showing the extensive hydrogen bonding.

be filtered by the glomeruli and excreted by the kidneys, giving rise to pigmented urine. A small fraction is reabsorbed by the renal tubules.

The significance of hyperbilirubinaemia

Less than 500 mg of bilirubin is produced each day, but the normal liver can conjugate up to 1500 mg/day. This large functional reserve is one reason why the plasma bilirubin concentration is an insensitive test for liver disease, since it begins to rise only after significant liver damage has occurred. In more than 95% of the apparently healthy adult population, the plasma bilirubin concentration is <25 μmol/L and this can be considered the upper limit of the reference range (Fig. 13.9). It is virtually all unconjugated. Jaundice can be detected by experienced observers when the plasma concentration exceeds 50 μmol/L and, when it rises above 100 μmol/L, it will be apparent to lay observers.

The absolute concentration of plasma bilirubin is of little help in establishing a diagnosis. Its measurement is, however, important in monitoring the progress of diseases such as primary biliary cirrhosis, where changes are of prognostic significance (see p. 248), neonatal jaundice (with respect to the need for exchange transfusion), for assessing response to treatment (such as surgical relief of bile duct obstruction) and to detect hyperbilirubinaemia that is suspected, but not clinically apparent. In addition, the absolute concentration of bilirubin is important in prescribing the correct dose of certain cytotoxic agents, especially if they are normally excreted in bile.

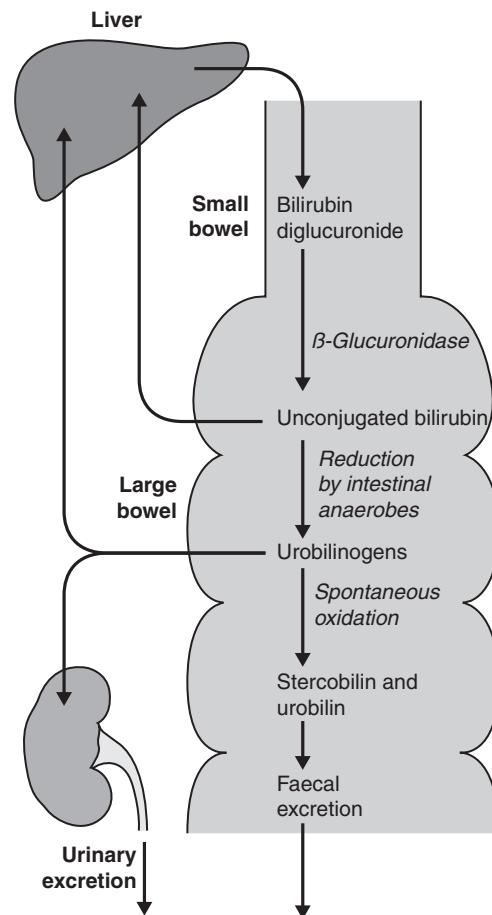


FIGURE 13.8 ■ Enterohepatic circulation of bilirubin and urobilinogen.

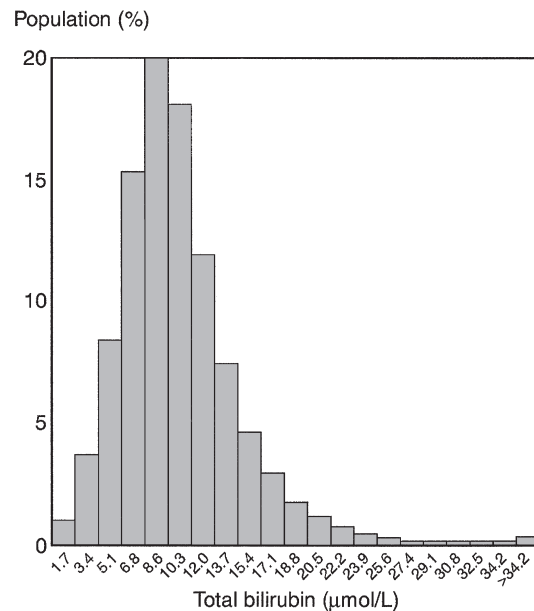


FIGURE 13.9 ■ Distribution of serum bilirubin concentrations in 18454 apparently healthy men. (From Bailey A, Robinson D, Dawson M 1977 Does Gilbert's disease exist? Lancet i:931-933, with permission).

For example, in patients treated with doxorubicin for malignant liver disease, the dose must be decreased in the presence of hyperbilirubinaemia to avoid drug-induced myelosuppression. The highest concentrations of bilirubin (sometimes approaching 1000 $\mu\text{mol/L}$) are seen when conjugated hyperbilirubinaemia of any cause is complicated by renal failure.

Tests for quantitation of bilirubin and its conjugated and unconjugated fractions

Current methods for the determination of plasma bilirubin are based on the diazo coupling of the pigments first described by Ehrlich in 1883. Two molecules of azodipyrrrole, which can be measured spectrophotometrically, are formed per molecule of bilirubin. In 1916, van den Berg and Muller noted that bilirubin from patients with obstructive jaundice reacted 'directly', whereas bilirubin from patients with haemolytic jaundice reacted 'indirectly', that is, an accelerator such as alcohol was required. The direct reacting pigment was later identified as conjugated bilirubin (CB) and the indirect reacting pigment as unconjugated bilirubin (UCB).

Although the distinction was found to be useful clinically, it soon became apparent that there is not, in fact, a precise relationship between indirectly reacting bilirubin and UCB, and directly reacting bilirubin and CB, respectively. In particular, direct measurements overestimate CB at low concentrations and underestimate it at high concentrations. To overcome these limitations, the alkaline methanolysis method was developed. The bilirubin mono- and diglucuronide conjugates are converted to the corresponding mono- and dimethyl esters by treatment with alkaline methanol. Unconjugated bilirubin is not affected by the reaction and is extracted into chloroform with the methyl ester derivatives. The pigments can then be separated and quantified by high performance liquid chromatography (HPLC). This approach shows that CB is virtually undetectable in plasma from healthy subjects or from patients with Gilbert syndrome. Techniques permitting accurate estimation of the two fractions at low bilirubin concentrations may become available routinely, and the detection of CB might then become the most specific and sensitive test of hepatobiliary function. Bilirubin bound to albumin (bili-albumin, delta bilirubin), probably covalently, is a third form of bilirubin. It may account for up to 90% of total bilirubin in both hepatocellular and cholestatic jaundice, although both impaired excretion and an intact conjugation mechanism are required for its formation. It is not detectable in healthy subjects or patients with unconjugated hyperbilirubinaemias, including Gilbert syndrome. It is this form of bilirubin that persists in the plasma of patients recovering from jaundice after bilirubin has ceased to be detectable in the urine. The precise mechanism by which jaundice occurs in patients with severe, long-term liver dysfunction has still not been firmly established. Both conjugated and unconjugated bilirubin are present, but the former predominates. Clinically, most patients retain the ability to conjugate bilirubin but have pigmented, bilirubin-containing urine. It is likely that conjugated bilirubin refluxes

retrogradely across the sinusoidal poles of hepatocytes because excretion via the biliary canalicular membrane becomes rate-limiting in the severely damaged liver. Leakage from damaged canaliculi via the paracellular pathway may also occur.

Plasma enzyme activities

As noted above, the enzymes used as liver tests, have little or no physiological function in the plasma: rather, they are empirical markers of some type of liver damage.

Alkaline phosphatase

The term alkaline phosphatase (ALP) describes a group of enzymes that hydrolyse phosphate esters at alkaline pH, although the physiological substrate within the body is not known. Alkaline phosphatase is present in several tissues, but the plasma component comprises mainly contributions from liver, bone, intestine and (during pregnancy) the placenta. There are two liver-derived isoenzymes. One derives from the hepatocyte and the other from the exterior surface of the biliary canalicular membrane. The latter probably enters the bloodstream via the paracellular pathway (see Fig. 13.3), although, in the presence of obstruction, it becomes distributed throughout the cytoplasm and may enter the plasma directly from across the plasma membrane. The function of hepatic ALP is unknown, but it may be involved in the transport of bile acids into bile.

The activity of plasma ALP rises when there is any form of obstruction to the biliary tract, whether mechanical or otherwise. The original assumption that this was due to failure of clearance by the obstructed liver was shown in the 1960s to be false. It was found that the clearance of ALP is not dependent on a patent biliary system and that during cholestasis only the liver ALP increases in activity, not the bone and intestinal isoenzymes. Subsequent studies have shown that this increase is from *de novo* synthesis of alkaline phosphatase, and it is possible in rat models to show that this is due to increased translation of the ALP mRNA rather than to increased transcription. The activity is definitely of biliary origin and the rise can be inhibited by blocking RNA synthesis. Biliary tract obstruction on its own is not sufficient to give rise to increased plasma activity of liver ALP; it appears that a high concentration of bile acids is also required, perhaps to solubilize the ALP bound to the canalicular membrane.

In cholestatic diseases due to biliary tract obstruction, the site of the obstruction may be at any level, from high in the small intrahepatic bile ducts (e.g. in primary biliary cirrhosis) down to the common bile duct (e.g. with gallstones). A rise in ALP activity usually predates the onset of clinical jaundice and, in those situations where surgical relief is possible, the return of plasma bilirubin to within the reference range usually precedes that of ALP. In a patient with symptoms of pain over the liver, an elevated ALP activity in the presence of a normal plasma bilirubin concentration is strongly suggestive of a hepatic space-occupying lesion, such as an intrahepatic tumour or an infiltrative disorder.

Overcoming the lack of tissue specificity. Increases in the activity of plasma ALP are not specific for liver disease. Modest increases occur in pregnancy, during periods of rapid bone growth in childhood and adolescence, and because of disease at other sites, particularly bone disease in which there is increased osteoblastic activity. While bone disease (hepatic osteodystrophy) may be a complication of longstanding cholestatic liver disease, the predominant form is osteoporosis, not osteomalacia, and the increased ALP activity is not due to the coexistent bone disease.

The specificity of the test can be enhanced by measuring specific ALP isoenzymes. Early studies used starch or polyacrylamide gel electrophoresis on the serum sample followed by specific staining of the gels for ALP; agarose gel electrophoresis is now more commonly used. In normal serum, two or three distinct bands can usually be identified, corresponding to the liver, bone and intestinal ALP isoenzymes. The liver band moves most rapidly towards the anode, with a more diffuse bone band closely behind. When present, the intestinal band lies behind the bone band. Additional bands are seen during pregnancy (the same isoenzyme is also detected occasionally in patients with malignancy, and is known as the Regan isoenzyme). In subjects with hepatocellular carcinoma, occasionally there is an additional, fast-running, band. The technique is semiquantitative, but it is usually visually obvious which of the isoenzymes is responsible for the increase in total activity.

An alternative approach is to repeat the standard ALP assay after first heating the serum at 56°C for 15 min. The liver and bone isoenzymes are sensitive to this treatment, and if the increased total activity is owing to either one of these, it will be reduced to about 40% or 15% of the original value, respectively. The placental isoenzyme activity remains unaffected by heating. This method is not as discriminatory as electrophoretic separation of the isoenzymes, but is easier to perform and can be used as a rough guide.

In practice, however, when there is doubt about the origin of an increased plasma ALP activity, it is customary to examine the results in relation to other enzymes, elevations in the activities of which are more liver specific. Thus, if the γ -glutamyltransferase (γ GT, see below) is also elevated, it may be inferred that the increase in ALP activity is probably of hepatic origin. In a few situations, elevation of ALP activity, apparently with hepatic specificity, may be seen in patients without overt liver disease, particularly those with Hodgkin lymphoma and some infections.

One specific situation in which isoenzyme analysis is of particular benefit is in patients with benign transient hyperphosphatasaemia (see Fig 13.10). First identified in children, it is now recognized also in adults. Serum ALP activity is typically >1000 U/L, with normal plasma activities of the aminotransferases and γ GT. It is usually associated with recent or intercurrent infection, often of the gastrointestinal tract. The mechanism is thought to be removal of sialic acid residues from circulating ALP by bacterial endotoxin or other substances released by the infecting organism. This alters the recognition of ALP by

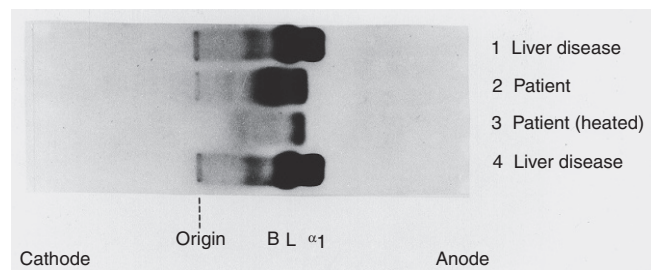


FIGURE 13.10 ■ Alkaline phosphatase isoenzymes in benign transient hyperphosphatasaemia. Separation of plasma ALP isoenzymes on cellulose acetate membrane. The positions of bone (B), liver (L) and biliary (α_1) isoenzymes are indicated. The electrophoretic pattern and heat stability of transient hyperphosphatasaemia are illustrated in lanes 2 and 3. (Courtesy of Dr S. B. Rosalki).

clearance receptors, prolonging its half-life in the circulation. The isoenzyme pattern is characteristic, showing the presence of two bands for each of the liver and bone isoforms. Treatment of the sample with neuraminidase, which removes terminal sialic acids from the carbohydrate side-chains of ALP, results in a pattern identical to that seen in a normal sample treated with neuraminidase. The importance of recognizing benign transient hyperphosphatasaemia is that invasive investigations such as endoscopic retrograde cholangiopancreatography are unnecessary.

The aminotransferases

These enzymes, previously designated (and still frequently referred to) as 'transaminases', catalyse the transfer of an amino group from an α -amino acid to an α -oxo acid. This is their intracellular function. An increase in their activity in the plasma is simply a marker of hepatocyte disruption and, presumably, has no useful function. The two most widely measured for clinical purposes are aspartate aminotransferase (AST) and alanine aminotransferase (ALT) both of which have a wide tissue distribution. Aspartate aminotransferase is found in liver, heart, skeletal muscle, kidney, brain, erythrocytes and lung. Alanine aminotransferase has a similar tissue distribution, but activities are much lower in extrahepatic tissue so that an increase in plasma ALT is more specific for liver disease. In view of this wide tissue distribution, it is not surprising that increased activities are seen in extrahepatic disease such as rhabdomyolysis, where the enzymes are released from damaged muscle cells. However, distinction between increased aminotransferase activities due to liver disease and those due to skeletal or cardiac muscle damage can usually be made by measurement of creatine kinase or the cardiac-specific troponins, respectively.

The increased plasma activities of these enzymes seen in patients with liver disease are presumed to originate from necrotic or damaged hepatocytes, although actual necrosis is not a prerequisite. It should not be assumed that the enzyme content of hepatocytes in patients with liver disease is necessarily the same as in normal hepatocytes; in patients with metastatic liver disease, hepatocyte enzyme concentrations may be several times higher than normal.

The aminotransferases are sensitive tests of hepatocellular dysfunction. This is particularly well demonstrated by the observation that during epidemics of viral hepatitis, aminotransferase activities may be markedly elevated in asymptomatic individuals with subclinical hepatitis. Generally speaking, the specificity of the test increases with enzyme activity. Activities above ten times the upper limit of the reference range are most frequently due to primary hepatocyte damage, so that the hepatic pathology will be some form of acute (viral or drug-induced) or chronic (such as autoimmune) hepatitis. Occasionally, acute cardiac failure or hypovolaemic shock may cause AST activities within this range, presumably from a combination of increased hepatic venous pressure, low cardiac output and arterial hypoxia. Exceptionally, such elevated activities may also be found in obstructive jaundice, particularly when there is acute biliary tract obstruction or where cholangitis supervenes on biliary tract obstruction.

Values below ten times the upper limit of the reference range are non-specific and no aetiological inference can be drawn. Importantly, a plasma ALT or AST activity within the reference range does not necessarily exclude significant hepatocellular damage, as assessed histologically in some forms of chronic liver disease (e.g. autoimmune hepatitis). In several liver disorders, more discrimination can be gained by considering the aminotransferase activities in relation to the ALP or to each other. In general, the higher the AST/ALP ratio, the more likely the underlying condition is to be some form of hepatitis. Conversely, the lower this ratio, the more it is indicative of a cholestatic disorder. Plasma ALT activity tends to be only modestly raised in patients with alcoholic hepatitis compared with other forms of hepatitis, whereas the AST activity is often more markedly elevated. Thus, an AST/ALT ratio of >2 , in a patient who appears on clinical grounds to have a hepatic illness, strongly suggests that alcohol is involved.

The mitochondrial isoenzyme of AST (mAST). There are two forms of AST, one cytosolic and the other mitochondrial (mAST). The latter is synthesized under control of nuclear DNA as a precursor (pre-mAST), which is rapidly transferred across the mitochondrial membrane and then converted into mature mAST. The mitochondrial isoenzyme of AST accounts for about 80% of total AST activity within the liver cells. With the development of immunochemical methods for measurement of this enzyme in serum, there has been considerable interest in the use of the mAST/total AST ratio as a marker of chronic alcohol consumption. While preliminary studies suggested that this test adequately distinguishes between those who consume excess alcohol and healthy subjects, irrespective of the presence or absence of liver disease, and is only elevated in association with chronic abuse, this test has not been adopted into routine laboratory practice.

γ -Glutamyltransferase

This is a microsomal enzyme responsible for transfer of glutamyl groups from γ -glutamyl peptides to other peptides or amino acids. It may be involved in the transport of peptides across cell membranes as γ -glutamyl peptides.

Although widely distributed throughout most body organs except muscle, the plasma activity is mainly attributable to the liver isoenzyme. γ -Glutamyltransferase has poor specificity for liver disease and, especially in a patient with jaundice, it adds little to the information gained from measurement of AST and ALP. However, its measurement can be useful in two particular circumstances. First, when the origin of an elevated serum ALP is uncertain, a concomitant elevation in γ GT suggests that the ALP is of hepatic origin. The second relates to the controversial area of the relationship of γ GT to chronic alcohol consumption.

The laboratory tests for chronic alcoholism are discussed in the next chapter. Here, we attempt to summarize what is accepted about γ GT and excessive alcohol consumption. In those who consume excess alcohol and who have chronic liver disease, the absolute plasma activity of γ GT is higher than in those without significant liver disease, and there is a tendency for activities to remain high after abstinence. This suggests that the elevated plasma γ GT activities in such individuals may be due to induction of the enzyme by the alcohol and/or because of the liver damage. In contrast, among alcoholic patients without liver disease, only about half will have a raised γ GT and this will usually fall to normal after eight weeks of abstinence. The extent of the increase in γ GT activity above normal is not directly related to either the amount of alcohol consumed or the duration of its consumption. It is evident from the above that the efficiency of using γ GT for screening populations for excessive alcohol consumption will be poor. False positive results occur in those taking enzyme-inducing drugs (including some herbal remedies) and false negative results occur in those who do not have liver disease. Nonetheless, the finding of a markedly raised γ GT (greater than five times the upper limit of the reference range) provides good reason to enquire diligently about possible alcohol abuse. γ -Glutamyltransferase remains the best of the simple laboratory screening tests and, depending on the population studied, the sensitivity is in the order of 50% and the specificity is about 85%.

Glutathione S-transferase (GST)

Isoenzymes of GST are involved in the detoxification and conjugation of several electrophilic compounds with glutathione. As noted above, bilirubin and bile acids bind strongly to GST, which is also involved in the metabolism of endogenous compounds such as some of the prostaglandins. Various isoenzymes exist; virtually all the plasma GST- α activity is of hepatic origin and its measurement seems to be a sensitive test of acute hepatocellular damage. It has a half-life in blood of only about 90 min and is, therefore, a rapidly responding marker of liver injury. However, the currently available immunoassay does not lend itself to large-scale routine screening and is therefore not widely used in clinical practice.

Plasma proteins

The concentrations of proteins in plasma reflect the balance between availability of their precursors and their

rates of synthesis, release and clearance, as well as their volumes of distribution. It is, therefore, not surprising that in patients with liver disease, the concentrations are very variable. Measurements of some plasma proteins (e.g. albumin, coagulation proteins) can provide useful information about general hepatic function; measurements of others (e.g. α_1 -antitrypsin, caeruloplasmin) are used in the diagnosis of specific conditions affecting the liver.

Albumin

This is the major protein in plasma and it is synthesized exclusively by the liver. The liver produces about 12 g of albumin each day; of the total body pool of 300 g, about 60% is in the extravascular pool and 40% in the intravascular pool. The plasma half-life is about 21 days. Albumin is responsible for maintaining plasma oncotic pressure and binds several hormones, anions, drugs and fatty acids. There is no doubt that as chronic liver disease progresses, the plasma albumin concentration tends to fall and, in this context, it is a major prognostic factor (see p. 248). Nonetheless, it must not be assumed that plasma albumin concentration is simply an indicator of hepatic synthetic function.

The rate of hepatic albumin synthesis falls in the face of inadequate protein intake. This is a frequent occurrence in patients with advanced liver disease, and particularly those in whom excessive alcohol consumption is implicated. However, even when the rate of synthesis falls, plasma concentrations may remain within the reference range because of a compensatory reduction in the rate of degradation. Furthermore, hypoalbuminaemia may occur in the face of normal, or even increased, rates of synthesis when protein 'leaks' into lymph, ascites or otherwise into the extravascular compartment. Hypoalbuminaemia may also be the result of renal loss of albumin in the nephrotic syndrome, gastrointestinal tract loss in protein losing enteropathy or loss through the skin in burns patients.

Prothrombin time and coagulation factors

The prothrombin time (PT) is a test that has become widely used in hepatology. Quick's one-step prothrombin time measures the rate at which prothrombin is converted to thrombin in the presence of thromboplastin, calcium, fibrinogen and other coagulation factors (V, VII and X). In turn, the thrombin leads to the conversion of fibrinogen to fibrin. Prothrombin and factors VII, IX and X all require vitamin K to become active. In the UK, the use of the PT has been widely replaced by the use of the international normalized ratio (INR). This is derived by dividing the PT of the patient by that of the control. Thus, healthy individuals will have an INR of 1 (<1.2 is the accepted normal value), while a patient with a PT of 120s where the control is 12s would have an INR of 10 (see Chapter 26).

There are two reasons why patients with liver disease may have a prolonged PT, each with different diagnostic implications. First, the liver may be so damaged that it cannot adequately synthesize the clotting factors that require vitamin K for their activation. The half-life of

factor VII is about 6h, so that a prolonged INR is a reliable indicator of the presence of acute liver failure caused by a recent, overwhelming insult to the liver. Second, since vitamin K is a fat-soluble vitamin, it may be deficient because of impaired fat absorption when there is obstructive jaundice. The latter situation is remediable by parenteral administration of vitamin K. Thus, return of the INR to the normal range within 18h may be taken as evidence of obstructive jaundice, whereas failure to respond implies severe parenchymal disease.

The concentration of factor V is becoming a widely-used test in assessing the severity of acute liver failure, particularly in continental Europe; low factor V concentration is associated with a poor prognosis.

α -Fetoprotein

Under normal circumstances, the synthesis of this protein, the fetal equivalent of albumin, virtually ceases shortly after birth. However, the protein is synthesized in large amounts (and becomes detectable in plasma) in about 70% of patients with hepatocellular carcinoma, and to lesser degrees in certain other tumours and benign liver diseases (see Chapter 42).

α_1 -Antitrypsin

This protein is the major α_1 -globulin and responsible for 90% of plasma tryptic inhibitory capacity. α_1 -Antitrypsin deficiency is a major cause of chronic liver disease in children (see Chapter 14) and, less commonly, of chronic liver disease presenting in adulthood.

Transferrin

Transferrin is the major circulating iron-binding protein, and its concentration correlates with the total iron-binding capacity of plasma. In states of iron overload, such as hereditary haemochromatosis (see Chapter 14), the concentration is modestly reduced, but the saturation is 55–100% compared with the reference range of 30–40%. Plasma transferrin has also been suggested to be a more appropriate overall indicator of protein synthesis by the liver in view of its shorter half-life, as compared with albumin.

Transferrin is a glycoprotein that exists in the circulation in various forms, containing up to six terminal sialic acid residues on its carbohydrate side chains. In healthy individuals, the trisialo-, tetrasialo- and pentasialo- forms predominate. Alcohol appears to inhibit the glycosylation of several glycoproteins, including transferrin, and in subjects consuming excessive amounts of alcohol, plasma transferrin often lacks up to four of these sialic acid residues, resulting in asialo- and disialotransferrins, which are now collectively termed carbohydrate-deficient transferrin (CDT). Consumption of more than 80g of alcohol/day leads to an increase in the plasma CDT concentration, irrespective of any underlying liver disease. This returns to normal concentrations within two weeks of abstinence. Carbohydrate-deficient transferrin measurement has been proposed as a marker of excessive alcohol consumption, although it is not in widespread use; sensitivity is

approximately 80% and specificity somewhat higher, but results do not reflect the nature or severity of any liver disease (see Chapter 14).

Caeruloplasmin

This protein, the principal copper-containing protein in the plasma, has oxidase activity including the ferroxidase activity essential for the oxidation of Fe(II) to Fe(III). It consists of a single polypeptide chain containing six copper atoms, but there is minimal turnover of the copper, suggesting that it does not act as a copper transporter in the manner that transferrin is the circulating iron carrier. In the context of liver disease, it plays an important role in the diagnosis of Wilson disease since it is virtually absent from plasma in this condition.

Acute phase reactants

In response to tissue damage, there is an increase in hepatic synthesis of several plasma proteins, notably C-reactive protein, α_1 -antitrypsin, antichymotrypsin, fibrinogen and caeruloplasmin. These are an integral part of the inflammatory response, which is described in greater detail in Chapter 30.

Immunoglobulins

Measurement of immunoglobulins is useful in the diagnosis of several liver disorders and for monitoring response to therapy in some. Elevated plasma concentrations often indicate an underlying inflammatory process. An indication of elevation in the immunoglobulins can be obtained by subtracting the plasma albumin concentration from the total protein concentration. This gives a rough measure of the total globulin fraction, of which the immunoglobulins are the major component. Concentrations of the individual immunoglobulin types (IgG, IgA, IgM) can then be determined by various techniques. This can provide helpful information, because different liver disorders are associated with selective elevations in the concentrations of the three immunoglobulin types. Thus, the IgG fraction is usually markedly elevated in autoimmune hepatitis, IgM in primary biliary cirrhosis, and IgA in alcoholic liver disease. Repeated measurement of IgG concentration is also useful in monitoring response to therapy in autoimmune hepatitis. This topic is discussed in more detail in Chapter 30.

Bile acids

In healthy individuals, the concentrations of plasma bile acids (PBAs) are determined by the difference between the amount absorbed from the gut and that taken up by the liver. The latter is fairly constant and, normally, it is the amount absorbed from the gut that determines plasma concentrations. Under pathological conditions, hepatic blood flow and clearance become the determining factors. In this situation, measurement of PBA may be considered as an endogenous clearance test. Accumulation of bile acids in the skin is thought to be associated with the development of the pruritis that is a relatively common feature of cholestatic liver disease.

Measurement of total plasma bile acids using simple automated enzymatic assays has proved to be of value in the diagnosis and monitoring of intrahepatic cholestasis of pregnancy. Total PBA measurement has also been used in the investigation of pruritis in children and adults with cholestatic liver disease, although interference in the assay from ursodeoxycholic acid, which may be used in treating these conditions, has limited its application. Fractionated bile acid analysis using high performance liquid chromatography and mass spectrometry is needed for the diagnosis of the rare inborn errors of bile acid metabolism (e.g. oxosteroid 5 β -reductase deficiency and amidation defects). The role of bile acids in carbohydrate and lipid metabolism is currently a subject of considerable interest.

Quantitative evaluation of liver function

The very limited extent to which standard liver tests give any quantitative measurement of hepatic function has already been discussed. To overcome these deficiencies, several tests that measure the efficacy of hepatic clearance of certain compounds, exogenous or endogenous, have been investigated over the past 20–30 years. To date, none has entered routine clinical practice. The main reason is that they require administration of the test substance and then subsequent collection of blood or other body fluids, sometimes over a protracted period. Despite some simplification of these tests, the information gained is seldom seen as justifying the necessary expenditure of time and effort.

However, with the advent of liver transplantation as a realistic therapeutic option in many forms of liver disease, it is becoming increasingly important to obtain more precise prognoses based on an accurate assessment of liver function. Also, many liver diseases are of such chronicity that a test that could reliably assess the efficacy of new drugs, in terms of their ability to improve liver function, would provide a real advance in clinical management of these conditions.

Pharmacological basis and practical requirements of clearance tests

The concept of ‘clearance’ as a test of hepatic function is analogous to the more familiar use of clearance as an index of renal function. The substance chosen should be non-toxic, exclusively eliminated by the liver and easily measurable. The principle of these tests is as follows:

$$\begin{aligned} \text{clearance} &= \text{hepatic removal} / \text{arterial plasma concentration} \\ \text{hepatic removal} &= \text{plasma flow} \times (\text{arterial concentration} \\ &\quad - \text{hepatic venous plasma concentration}) \end{aligned}$$

Thus:

$$\text{clearance} = \frac{\text{plasma flow} \times (\text{arterial concentration} - \text{hepatic venous plasma concentration})}{\text{arterial plasma concentration}}$$

Clearance is thus a function of plasma (and hence blood) flow and the second term in the above equation,

known as the hepatic extraction ratio (E). When the extraction ratio approaches 1, the clearance equates to liver plasma flow. In healthy subjects, the value of E for indocyanine green is about 0.9 and the clearance of this substance is, indeed, a good indicator of blood flow. In liver disease, however, E may fall to as low as 0.2 and the clearance is no longer dependent on blood flow but more on the ability of the liver to clear the compound. Compounds such as aminopyrine, antipyrine and caffeine have extraction ratios of <0.25 in healthy subjects and, providing that saturating doses are given, clearance is largely independent of blood flow.

Indocyanine green is not conjugated in the liver, although its removal from the plasma depends on hepatic function. Ten milligrams per kilogram body weight is administered intravenously and samples taken at 3 min intervals between 3 and 15 min. The percentage disappearance rate (PDR) is calculated as:

$$\text{PDR} = \left(0.693 / t_{1/2}\right) \times 100$$

where $t_{1/2}$ is the half-life of indocyanine green. The reference range is 17–22%; values below this range are considered abnormal.

Other tests of liver function

Serum tests for hepatic fibrosis

Fibrosis, leading ultimately to cirrhosis, is a characteristic feature of progressive chronic liver disease and is responsible for much of the associated functional disturbance and portal hypertension. In chronic liver disease, collagen is laid down in the perisinusoidal space (see Fig. 13.3) and the space of Disse, to the extent that the fenestrations are lost and the sinusoids become capillarized with a true basement membrane. Hepatic stellate cells are likely to play a determining role in the fibrotic process and, increasingly, therapy is aimed at reversing, or at least inhibiting, this process. Until recently, the only method of assessing the extent of fibrosis or response to therapy was by histological examination of liver tissue but, because the fibrosis is often not uniformly distributed throughout the liver, there is a risk of sampling error. Transient elastography (Fibroscan[®]), a technique that measures liver stiffness, correlates with histological staging of fibrosis. A low amplitude shear wave is generated from a probe applied to the thoracic wall at the level of the right lobe of the liver. The velocity of propagation is directly proportional to liver stiffness. This estimates existing fibrosis but cannot provide any information on whether the fibrotic process is historical or ongoing. For this reason, biochemical markers of fibrogenesis have been sought for many years.

Several biochemical candidates have been proposed as markers of fibrosis either in isolation or combined together in a scoring system. Some of these incorporate standard biochemical tests whilst others involve measurement of compounds involved in the fibrogenic process (Table 13.1). The increasing prevalence of liver disease due to infection with the hepatitis C virus has provided focus to the search for non-invasive markers of clinically significant fibrosis.

TABLE 13.1 Scoring systems for assessing fibrogenesis

Score	Components
PGA index	Prothrombin time, γ GT, apolipoprotein A1 ($\pm \alpha_2$ -macroglobulin)
Fibro Test [®]	Haptoglobin, α_2 -macroglobulin, apolipoprotein A1, γ GT, bilirubin, (\pm ALT)
Fibrometer [™] test	Platelets, prothrombin time, AST, α_2 -macroglobulin, hyaluronate, urea, age
Hepa Score [®]	Bilirubin, γ GT, α_2 -macroglobulin, hyaluronate, age, sex
ELF score	PIIINP, hyaluronate, TIMP-1

ALT, alanine aminotransferase; AST, aspartate aminotransferase; PIIINP, amino-terminal procollagen type III; TIMP-1, Tissue inhibitor of metalloproteinase 1.

Collagen metabolites. Quantitatively, types I and III collagen are involved in hepatic fibrosis but type IV is also important in basement membranes. Measurement of the concentration of the amino-terminal procollagen type III (PIIINP) in plasma is a non-invasive marker of collagen type III metabolism. The PIIINP concentration reflects fibrinogenesis (i.e. the rate of ongoing fibrosis) rather than the absolute amount of fibrosis present in the liver. Nonetheless, PIIINP measurements are of value for indicating the development of fibrosis in patients receiving potentially hepatic fibrogenic drugs such as methotrexate. Other approaches, such as measurement of the carboxy-terminal cross-linking domain of type IV collagen, are being actively investigated.

Hyaluronate. Hyaluronate is a high molecular weight glycosaminoglycan with a structural role in the extracellular matrix. It is synthesized by hepatic stellate cells and cleared by specific uptake by endothelial cells. Increased plasma concentrations may be linked to endothelial dysfunction that occurs as fibrosis progresses.

Tissue inhibitor of metalloproteinase 1. Tissue inhibitor of metalloproteinase 1 (TIMP-1) is a member of the tissue inhibitors of metalloproteinases (MMP) family. These MMPs are secreted by fibroblasts and Kupffer cells and MMP-2 in particular is upregulated by type 1 collagen: MMP-2 is probably involved in the breakdown of normal extracellular matrix. Tissue inhibitor of metalloproteinase 1, which inhibits MMP-2, may also promote cell proliferation and have apoptotic properties. It is used in the calculation of the European liver fibrosis score (see below).

European liver fibrosis score (ELF). A combination of PIIINP, hyaluronate and TIMP-1 has been proposed as a better marker of fibrogenesis than the component parts on their own. The ELF score is produced from the equation below:

$$\text{ELF} = -7.412 + (\ln[\text{HA}] \times 0.681) + (\ln[\text{PIIINP}] \times 0.775) + (\ln[\text{TIMP}-1] \times 0.494)$$

All measurements are expressed as $\mu\text{g/L}$. Normal scores fall between -1.3 and $+0.3$.

The first major study of 1021 subjects with chronic liver disease of various causes showed a sensitivity of 90% and specificity of 92% for the presence of hepatic fibrosis. Further investigations into potential applications of the ELF score are ongoing.

USES OF LIVER FUNCTION TESTS

Differential diagnosis of jaundice

The differential diagnosis of jaundice is an exercise in applied physiology and anatomy. Provided the reader accepts that there is seldom a single mechanism, the traditional classification into pre-hepatic, hepatic and post-hepatic remains useful. In each case, bilirubin is being produced faster than it can be cleared.

Pre-hepatic jaundice

In pre-hepatic jaundice, an increased bilirubin load is presented to the liver and the plasma bilirubin concentration reaches a steady state at a higher concentration, typically in the range $50\text{--}75\ \mu\text{mol/L}$. A haemolytic process due either to fragile red cells (e.g. sickle cell disease, congenital spherocytosis) or an autoimmune attack on the red cells is usually implicated. The bilirubin is unconjugated and bound to albumin and thus does not appear in the urine, hence the term 'acholuric jaundice' is used to describe the jaundice accompanying haemolysis. Since there is no liver disease, the liver can conjugate and excrete the increased bilirubin load, but the amounts of urinary and faecal urobilinogen will be increased. As the haemoglobin concentration starts to fall as a result of the haemolytic process, the bone marrow will attempt to compensate by increasing the output of red cells. The maximum increase that the marrow can achieve (about eight times the normal rate) corresponds to a plasma bilirubin concentration of $\sim 75\ \mu\text{mol/L}$.

If, however, haemolysis occurs at such a rate that compensatory marrow hyperactivity cannot keep pace, then the haemoglobin concentration will fall and higher plasma bilirubin concentrations may be seen. On the other side of the equation, if the liver is damaged, then its ability to take up the bilirubin will be limited and concentrations of bilirubin $>75\ \mu\text{mol/L}$ may occur. Thus, when a patient is found to have mild jaundice and there is no bilirubin in the urine, examination of the haematological indices will usually be sufficient to make the diagnosis of some form of haemolytic anaemia, especially as other liver tests will, by contrast, be normal. If there is no evidence of haemolysis, the cause will usually be Gilbert syndrome (see below). Estimation of the conjugated and unconjugated fractions becomes of particular importance in determining if, in a patient already known to have chronic liver disease, haemolysis is contributing to deepening jaundice.

Urinary bilirubin and urobilinogen. The presence of bilirubin in urine (usually detected by commercially available dip-sticks) is always pathological and indicative

of hepatic or biliary tract disease. Bilirubinuria may precede hyperbilirubinaemia and, therefore, be a very sensitive test of impending liver damage, for example in the early, presymptomatic stage of viral hepatitis.

If bile pigments reach the gut in reduced amounts (as in obstructive jaundice), the amount of urobilinogen in the urine will decrease accordingly. If no urobilinogen can be detected on repeated testing over several days, complete biliary tract obstruction, the most common cause of which is pancreatic carcinoma, may be assumed. Although urine urobilinogen can be assessed qualitatively using dipsticks, testing is now rarely performed as there are much more reliable and accurate ways of diagnosing biliary obstruction.

Hepatic (hepatocellular) jaundice

This type of jaundice occurs in conditions characterized by primary damage to the hepatocytes, that is, acute and chronic hepatitis of any cause. The excess bilirubin is mainly conjugated.

Post-hepatic (cholestatic) jaundice

Cholestasis means failure of biliary secretion. As originally defined, the term referred to failure at the level of the hepatocytes, but in clinical practice it is more broadly defined to include obstruction to bile flow occurring anywhere between the biliary canalicular membrane of the hepatocyte and the gut. In cholestatic jaundice, the hyperbilirubinaemia is due to conjugated bilirubin.

Further investigation. The crucial decision for the clinician to make in a patient with conjugated hyperbilirubinaemia is whether or not a surgical approach is appropriate, as misdiagnoses can have grave consequences. The role of the liver function tests is clear: it is to suggest the most appropriate second-line investigation. In general, if the plasma aminotransferases are elevated and there is only slight increase in ALP activity, the picture is regarded as 'hepatic'. The approach is primarily medical management, possibly guided by the result of a liver biopsy and other investigations. On the other hand, if the predominant rise is in ALP, the picture is regarded as 'cholestatic' and it is important to determine the level at which there is failure of bile flow. This is achieved using imaging techniques, for example ultrasound examination, endoscopic retrograde cholangiography or magnetic resonance cholangiography. Obstruction to bile flow outside the liver is usually an indication for surgical intervention.

Imaging techniques have become so widely available and reliable that the great majority of patients with jaundice of unknown origin will have at least an ultrasound or radiological examination irrespective of the liver function tests. Perhaps only in the primary care setting, if the clinical picture and 'hepatic' liver function tests suggest acute hepatitis, will imaging not be undertaken.

The inherited hyperbilirubinaemias

The inherited hyperbilirubinaemias, manifested mainly in childhood, are considered here because

they illustrate specific disturbances of the pathways of bile pigment metabolism. Other than the specific defect, the liver functions normally and, apart from the plasma bilirubin, the results of most of the standard biochemical liver tests are within their reference ranges. Conventionally, these disorders are classified as conjugated and unconjugated types. Unconjugated hyperbilirubinaemia is said to be present when $<25 \mu\text{mol/L}$ or $<20\%$ of an increased bilirubin concentration is in the conjugated form.

Unconjugated types

Crigler–Najjar syndrome. This is an extremely rare condition, caused by a complete absence (type 1) or considerable reduction (type 2) of UDP-glucuronosyl transferase activity with respect to bilirubin glucuronidation. It is inherited in an autosomal recessive manner. The condition shows considerable heterogeneity. There is a failure to conjugate bilirubin and the child becomes very deeply jaundiced within the first few days of life, most of the plasma bilirubin being in the unconjugated state. The unconjugated bilirubin can cross the blood–brain barrier and be deposited in the basal ganglia, leading to characteristic neurological abnormalities termed kernicterus, which can include extrapyramidal-type movement disorders, visual defects and hearing problems. In the most severe form (type 1), death usually ensues within the first year of life, although repeated phototherapy has been used successfully to reduce bilirubin concentrations and liver transplantation is now a viable option. In Crigler–Najjar syndrome type 2, the activity of UDP-glucuronosyl transferase can be increased by giving phenobarbital (which induces synthesis of the enzyme) in addition to phototherapy. Survival into adulthood is now well described.

Gilbert syndrome. This, in contrast to Crigler–Najjar syndrome, is a common (occurring in approximately 7% of the population) and entirely benign condition, characterized by recurrent episodes of mild jaundice, the plasma bilirubin concentration being $<100 \mu\text{mol/L}$ and the only abnormality in the standard liver tests. Jaundice tends to be more pronounced when the affected individual is tired, has an intercurrent illness (e.g. influenza) or has been fasting. It should be remembered that caloric deprivation often occurs during hospital admission for any reason, and the resultant jaundice in an individual with Gilbert syndrome may be regarded as indicating a significant hepatic complication if the diagnosis is not made.

Gilbert syndrome is caused by a reduction in the concentration of the isoform of UDP-glucuronosyltransferase, UGT-1A, responsible for bilirubin conjugation, owing to polymorphisms in the promoter region of the *UGT-1A* gene that reduces its expression. A two base-pair addition (TA) in the TA₍₆₎ TAA element of the *UGT-1A* promoter sequence, when present in the homozygous state (TA₍₇₎/TA₍₇₎), results in an inability to produce sufficient UGT-1A to respond to increases in demand for bilirubin conjugation. Genotyping for the defect is now available as a routine test for the syndrome, and has superseded other tests.

Conjugated types

Dubin–Johnson syndrome. This is an uncommon, benign condition inherited in an autosomal recessive manner and characterized by decreased biliary canalicular transport of conjugated bilirubin and organic anions. Dubin–Johnson syndrome is caused by mutations in *MRP2*, the gene encoding multidrug resistance-associated protein 2, and presents with mild fluctuating jaundice at any age from the neonatal period to adult life, usually with normal aminotransferase and ALP activities. It is often precipitated by use of the contraceptive pill or during pregnancy and, with the increasing availability of mutational analysis, is now frequently revealed during screening of family members with the condition. The liver biopsy shows black coarse inclusions of an as yet unidentified pigment.

A second characteristic finding in this syndrome is that most of the urinary coproporphyrin is type I, whereas in unaffected subjects it is mainly type III. The total amount of urinary coproporphyrin is normal and the abnormality is probably related to a differential excretory capacity for the two isomers.

Rotor syndrome. Like Dubin–Johnson syndrome, this is an autosomal recessive condition characterized by a moderate fluctuating jaundice, few symptoms and having a good prognosis. However, several differences from Dubin–Johnson syndrome are evident: in particular there is no pigment in the liver. The total urine coproporphyrin content is high but with a normal ratio of isomers. The gene mutation causing this syndrome has not yet been identified.

Monitoring response to therapy

Liver biopsy, sophisticated imaging tests and specific biochemical tests, outlined in Chapter 14, are helpful in establishing a precise diagnosis in liver disease, but they have limited utility in repeated appraisal of disease progression or monitoring response to treatment over short periods. This reflects both their cost (in the case of imaging) and safety aspects (in the case of liver biopsy). Herein lies the major role of the standard liver function tests.

There are three broad situations in which frequent monitoring is useful.

1. **Documenting the return of ‘normality’ in acute conditions and early detection of progression to chronicity.** For example, following an acute attack of viral hepatitis B or C, a variable proportion of patients will progress to a chronic carrier state, sometimes with progressive liver disease ending in cirrhosis and perhaps liver cell cancer. The chronic state is defined as the presence of signs of liver disease and/or abnormal liver tests six months after the onset of the acute illness, and heralds the need for therapeutic intervention. Monitoring of patients with hepatitis A is usually not necessary unless the attack is particularly severe. Under such circumstances, progressive deterioration in liver tests and prolongation of the prothrombin time may herald the onset of acute liver failure and necessitate referral to a specialist unit.

2. **Monitoring response to treatment of chronic conditions.** Good examples here are in the immunosuppressive treatment of autoimmune hepatitis and antiviral therapy in chronic viral hepatitis; other examples are described in Chapter 14. Specifically, monitoring of changes in aminotransferase activities is a useful (and cheap) means of assessing response in autoimmune hepatitis (along with measurement of plasma IgG concentration), and as an adjunct to detection of the reduction in (or clearance of) viral genomic material from the blood in chronic viral hepatitis.
3. **Detecting hepatotoxicity caused by drugs.** Monitoring of liver function tests is undertaken for all drugs during clinical trials prior to licensing; any abnormalities are taken as evidence of hepatotoxicity (see Chapter 14). In some instances, where a drug is licensed, but an acceptable risk of toxicity (mild and reversible) has been identified, monitoring may be recommended as a part of continuing surveillance.

Neonatal jaundice

Healthy full-term babies may be jaundiced from the 2nd to the 8th day of life, and premature babies from the 2nd to the 10th, or even the 14th day. This so-called 'physiological' jaundice seldom exceeds 100 $\mu\text{mol/L}$ and is predominantly unconjugated, reflecting a combination of increased red cell destruction, decreased hepatic uptake and reabsorption from the gut. Jaundice detectable on the first day of life, or outside the range described above, is pathological, as is the finding of conjugated hyperbilirubinaemia at any time.

When the concentration of bilirubin is very high (>300 $\mu\text{mol/L}$ in full-term infants, but lower in premature babies), the fat-soluble unconjugated bilirubin may cross the blood-brain barrier and cause brain damage (kernicterus). While measurement of plasma bilirubin is not indicated as routine in jaundiced babies, it does become important when the jaundice is deepening so that treatment by phototherapy and/or exchange transfusion can be instituted and kernicterus prevented.

The neonatal hepatitis syndrome

This syndrome, of which there are several causes, describes a neonate usually less than one month old with pale stools, dark urine and hepatomegaly, who is failing to thrive. In common with adult practice, the major problem is to identify those patients that are amenable to effective surgical intervention, particularly those having congenital biliary atresia (see Chapter 25). It is crucial to confirm that the biliary tract is atretic before embarking on surgery.

Tests of bile duct patency. The Rose Bengal dye test has proved useful in the past for identifying patients with extrahepatic neonatal obstruction. If <10% of an intravenously administered dose is excreted in the stools over a 72 h period, this is strong evidence for biliary obstruction. Currently, technetium-labelled hepatobiliary imaging agents such as hydroxyindolediacetic acid (HIDA) are more often used than Rose Bengal. In combination with liver biopsy, accurate results are achieved in the great majority of cases.

ABNORMAL LIVER FUNCTION TESTS IN ASYMPTOMATIC PATIENTS

With the use of large automated analysers, health screening programmes and medical examinations for insurance purposes, clinical biochemists are increasingly consulted about the significance of abnormal liver tests in patients who are entirely asymptomatic. While the potential for generating needless anxiety is great and not all conditions diagnosed are treatable, much evidence suggests that such abnormalities should be taken seriously, and that referral for a specialist opinion for further investigation is often indicated and justifiable.

Bilirubin

Isolated increases in plasma bilirubin usually indicate Gilbert syndrome. The concentration is <100 $\mu\text{mol/L}$ and usually in the range 20–50 $\mu\text{mol/L}$. This hyperbilirubinaemia is unconjugated and other standard liver function tests are normal (see p. 246). Measurement of haemoglobin, reticulocyte count and, if necessary, haptoglobin will eliminate haemolysis as a cause. Genetic testing is available.

Alkaline phosphatase

In young people, an isolated increase in the activity of this enzyme is often physiological. Increased activity occurs during periods of rapid bone growth and during pregnancy. In older females, the finding of isolated raised ALP activity of hepatic origin may be the first manifestation of primary biliary cirrhosis and estimation of antimitochondrial antibody (see Chapter 14) will speed the diagnostic process. Alternatively (more commonly in males), it may be suggestive of primary sclerosing cholangitis. Subclinical malignant liver disease (primary or secondary), usually found to be a space-occupying mass within the liver, may also produce an isolated rise in plasma alkaline phosphatase activity. Benign transient hyperphosphataemia (see p. 240) is an important condition to recognize, as ALP will return to normal over a 6–8-week period after any infection has cleared and invasive investigations can be avoided.

Aminotransferases

In about 60% of patients, isolated increases in the activities of these enzymes will be attributable to fatty infiltration of the liver associated with obesity, diabetes mellitus or excessive alcohol consumption. However, they may also be due to certain drugs and herbal remedies. Once coexistent drug administration has been excluded, most remaining individuals who have persistently elevated enzyme activities will prove to have some form of chronic hepatitis. Many of these, particularly subjects born outside the UK or those with a history of drug abuse, may have occult chronic viral hepatitis (B or C) – although subclinical hepatitis A is also a possibility. Others will have elevated plasma immunoglobulin concentrations and high plasma titres of autoantibodies (see Chapter 14); on histological examination, cirrhosis will often be present. A small proportion will have other

conditions such as haemochromatosis, Wilson disease or α_1 -antitrypsin deficiency.

γ -Glutamyltransferase

Marked and persistent elevation in γ GT activity may be an indicator of unrecognized excessive alcohol consumption, but it is by no means specific and may also be found in most subclinical liver diseases, or be a result of induction of the enzyme by a number of drugs, components of herbal remedies or environmental agents. Some non-hepatic conditions, including pancreatic disease, hypothyroidism and certain neurological disorders, may also give rise to isolated increases in γ GT activity. The increasing prevalence of obesity, type 2 diabetes mellitus and non-alcoholic fatty liver disease (NAFLD) has meant that an isolated increase in γ GT is most commonly a consequence of fat deposition in the liver.

NORMAL LIVER FUNCTION TESTS IN THE PRESENCE OF OVERT LIVER DISEASE

Standard liver function tests within the reference range do not exclude chronic liver disease. Entirely normal test results may be present in patients with 'inactive' cirrhosis, that is, where there is little or no ongoing liver damage or inflammation. Common examples include alcoholic cirrhosis after abstinence from drinking and autoimmune hepatitis that is well controlled on immunosuppressive therapy. In the presence of normal standard liver function tests, signs of portal hypertension (such as oesophageal varices, which are usually indicators of chronic liver disease) should also raise the possibility of portal vein thrombosis. In this situation, the INR is often modestly prolonged.

ROLE OF LIVER FUNCTION TESTS IN ASSESSING PROGNOSIS

Orthotopic liver transplantation for liver failure is now a well-established form of management in severe acute and chronic liver disease. Nonetheless, the operation is not without risk and precise timing of the procedure is still difficult. Ideally, it should be undertaken only when the patient's life expectancy is less than what might be expected after the operation. Several models that aim to permit calculation of likely survival without transplantation

have been developed. Some 'standard' liver function tests figure prominently among the variables considered in these models.

Chronic liver disease

An early system, still widely applied, is the Child–Turcotte classification, which was later modified by Pugh et al. (Table 13.2). This was an entirely empirical formula designed primarily to assess the degree of risk of liver failure in patients undergoing surgical operations, but has since been used as a general measure of severity of liver disease and hence likely survival. The Child–Turcotte–Pugh (CTP) classification has been criticized on the grounds that two of the measures (ascites and encephalopathy) are subjective, and that the patient is assigned one of three scores that do not reflect a spectrum of overall impairment of liver function and hence the need for transplantation. In addition, a 'ceiling effect' occurs when patients with laboratory values greater than the defined upper limits are not scored as being more ill than those falling within these limits. Furthermore, the CTP classification does not take into account the terminal stages of liver impairment when renal failure supervenes, a complication that indicates a very poor outcome.

The need for more robust and objective criteria on which to base selection for liver transplantation has prompted the recent development of a Model for End Stage Liver Disease (MELD) scoring system. This prognostic model for cirrhosis draws on the analysis of prospective datasets using logistic regression analysis and receiver operating characteristics, and uses serum total bilirubin and creatinine concentrations and INR.

The equation used to calculate the MELD score is:

$$R = 9.6 \times \ln(88.4 \times \text{creatinine in } \mu\text{mol/L}) + 3.8 \\ \times \ln(17.1 \times \text{total bilirubin in } \mu\text{mol/L}) + 11.2 \\ \times \ln(\text{INR}) + 6.43$$

MELD scores derived from many patients in a wide range of institutions dealing with the selection of patients for liver transplantation, fall within a range of values that vary from 6 (less ill) to 40 (gravely ill). The scores are used to decide whether an individual needs an urgent liver transplant within the next three months, with patients that have a MELD score of >16 being prioritized for transplantation.

TABLE 13.2 Pugh's modification of Child's classification for assessing the severity of liver disease

Clinical and biochemical measurements	1	2	3
Ascites	Absent	Slight	Moderate
Encephalopathy (grade)	None	1 and 2	3 and 4
Bilirubin ($\mu\text{mol/L}$)	<25	25–40	>40
Albumin (g/L)	>35	28–35	<28
International normalized ratio (INR)	<1.7	1.71–2.3	>2.3

The disease is considered mild in patients scoring 5 or 6 points (Grade A), moderate in those with 7–9 points (Grade B) and severe in those with 10–15 points (Grade C).

(From Pugh R N H, Murray-Lyon I M, Dawson J L et al. 1973 Transection of the oesophagus for bleeding oesophageal varices. *British Journal of Surgery* 60:646–50, with permission).

Acute liver failure

Liver transplantation for acute liver failure is now routinely performed but, again, laboratory investigations are important in assessing the likelihood of survival without transplantation. For example, in patients with paracetamol-induced acute liver failure, an arterial hydrogen ion concentration of $>50\text{nmol/L}$ ($\text{pH} <7.3$) or, if this feature is absent, the combination of grade 3 encephalopathy, an INR >6.5 and serum creatinine of $>300\mu\text{mol/L}$, indicate a very slight chance of spontaneous recovery and represent a strong indication to proceed to transplantation.

CONCLUSION

There are three categories of liver function tests. These are: the standard tests, for example plasma bilirubin and albumin concentrations and the activities of various enzymes; clearance tests, and tests used in the management of specific liver diseases.

In general, the standard liver function tests are poor indicators of hepatic function and seldom provide a specific diagnosis; it is important not to over-interpret their results. They may, however, provide a guide to further investigations that may give a specific diagnosis, for example imaging by isotopes, ultrasound or radiography or histological examination of liver biopsy material. The standard liver function tests can be of value prognostically and in screening for liver disease, and are vital in the monitoring of liver disease and its response to treatment.

Tests for specific liver disease, for example plasma α -fetoprotein, copper and caeruloplasmin, can be diagnostic and are also used in management.

Further reading

- Clermont RJ, Chalmers TC. The transaminase tests in liver disease. *Medicine* 1967;46:197–207.
- The classic paper on the subject of aminotransferases as a test of liver function.*
- Ferraris R, Colombatti G, Fiorentini MT et al. Diagnostic value of serum bile acids and routine liver function tests in hepatobiliary diseases. Sensitivity, specificity and predictive value. *Dig Dis Sci* 1983;28:129–36.
- Gressner AM, Gao CF, Gressner OA. Non-invasive biomarkers for monitoring the fibrogenic process in liver: a short survey. *World J Gastroenterol* 2009;15:2433–40.
- Hannuksela ML, Liisanantti MK, Nissinen AE et al. Biochemical markers of alcoholism. *Clin Chem Lab Med* 2007;45:953–61.
- Hulcrantz R, Glaumann H, Lindberg G et al. Liver investigation in 149 asymptomatic patients with moderately elevated activities of serum aminotransferases. *Scand J Gastroenterol* 1986;21:109–13.
- An important study discussing the significance of liver function test abnormalities detected in asymptomatic individuals.*
- Kamath PS, Weisner RH, Malinchoc M et al. A model to predict survival in patients with end-stage liver disease. *Hepatology* 2001;33:464–70.
- Kaplan MM. Serum alkaline phosphatase – another piece is added to the puzzle. *Hepatology* 1986;6:526–8.
- Describes the history of the use of serum alkaline phosphatase.*
- Kunde SS, Lazenby AJ, Clements RH et al. Spectrum of NAFLD and diagnostic implications of the proposed new normal range for serum ALT in obese women. *Hepatology* 2005;42:650–6.
- Moussavian SM, Becker RC, Piedmeyer JL et al. Serum gamma-glutamyl transpeptidase and chronic alcoholism: influence of alcohol ingestion and liver disease. *Dig Dis Sci* 1985;30:211–4.
- O'Grady J, Alexander GJM, Hayllar K et al. Early indicators of prognosis in fulminant hepatic failure. *Gastroenterology* 1989;97:439–45.
- Use of liver tests to assess prognosis in fulminant hepatic failure.*
- Penn R, Worthington DJ. Is gamma-glutamyltransferase a misleading test? *Br Med J* 1983;286:531–4.
- Pugh RNH, Murray-Lyon IM, Dawson JL et al. Transection of the oesophagus for bleeding oesophageal varices. *Br J Surg* 1973;60:646–50.
- Description of the Pugh modification of Child's classification of the severity of cirrhosis.*
- Rappaport AM, Borowy ZJ, Laugheed WM et al. Subdivision of hexagonal liver lobules into a structural and functional unit: role in hepatic physiology and pathology. *Anat Rec* 1954;119:11–34.
- Classic paper on the relationship between structure and function in the liver.*
- Trauner M, Boyer JL. Bile salt transporters: molecular characterization, function, and regulation. *Physiol Rev* 2003;83:633–71.

Acute and chronic liver disease

Adrian Bomford • Roy A. Sherwood

CHAPTER OUTLINE

CLASSIFICATION OF LIVER DISEASE 250

ACUTE HEPATITIS AND ITS SEQUELAE 250

Differential diagnosis 251

Acute viral hepatitis 251

ACUTE LIVER FAILURE 253

Laboratory features 253

Laboratory criteria for liver transplantation 253

CHRONIC HEPATITIS 254

Differential diagnosis of chronic hepatitis 254

PRIMARY BILIARY CIRRHOSIS 256

PRIMARY SCLEROSING CHOLANGITIS 256

ALCOHOLIC LIVER DISEASE 257

Ethanol metabolism 257

Liver pathology in alcoholic liver disease 257

Use of laboratory tests in clinical practice 258

Non-alcoholic fatty liver disease 258

THE CONCEPT OF CIRRHOSIS 259

Hepatic encephalopathy 259

Vascular disturbances in cirrhosis 259

Ascites 259

Acute kidney injury 260

Sex hormones and their binding proteins 261

Glucose intolerance 263

DRUGS AND THE LIVER 263

NEOPLASTIC DISEASE OF THE LIVER AND BILIARY TRACT 264

Hepatocellular carcinoma and α -fetoprotein 264

PARENTERAL NUTRITION 264

BACTERIAL INFECTIONS 264

INHERITED METABOLIC DISORDERS INVOLVING THE LIVER 265

Iron overload and hereditary haemochromatosis 265

Wilson disease 266

α_1 -Antitrypsin deficiency 268

The hepatic porphyrias 268

Cystic fibrosis 268

Other inherited metabolic diseases 269

LIVER TRANSPLANTATION 270

Preoperative assessment 271

The immediate postoperative period 271

Intermediate follow-up 271

Long-term monitoring 271

CONCLUSION 272

CLASSIFICATION OF LIVER DISEASE

Liver disease is best classified according to its presumed aetiology (Box 14.1). This is then qualified by the pathological state of the liver where known or inferred: usually cirrhosis, inflammation (i.e. hepatitis) or cholestasis. Thus, terms such as alcoholic cirrhosis, viral hepatitis or cholestasis of pregnancy all conform to this classification. Some idea of the severity of the disease is given by the terms 'compensated' and 'decompensated'. The term 'compensated' implies that, although the liver has been damaged it can, because of its large functional reserve, still function adequately. On the other hand, the term 'decompensated' implies that the liver is failing in some vital function. The terms are not precisely defined, but the development of ascites with or without peripheral

oedema, jaundice or hepatic encephalopathy, in a patient known to have liver disease, would all be regarded as signs of decompensation. Finally, the presumed evolution of the disease in time is given by the terms fulminant, acute, subacute or chronic. In chronic liver disease, additional terminology often employed to provide an indication of severity, especially when assessing histological changes in the liver, includes the terms active or inactive and mild, moderate or severe.

ACUTE HEPATITIS AND ITS SEQUELAE

The term 'hepatitis' indicates that there is hepatic inflammation, for which there are many causes: viral infections, drugs and toxins (including alcohol) are common, while

BOX 14.1 Classification of liver disease according to aetiology**Viral**

- Hepatitis A, B, C, D and E
- Epstein–Barr virus
- Cytomegalovirus
- Herpes simplex
- ‘Exotic viruses’

Genetic/metabolic

- Haemochromatosis
- Wilson disease
- Hereditary hyperbilirubinaemias
- α_1 -Antitrypsin deficiency
- Cystic fibrosis
- Hepatic porphyria
- Amyloid

Bacterial/spirochaetal

- Leptospirosis
- Tuberculosis
- Pyogenic liver abscess

Protozoal

- Kala-azar (visceral leishmaniasis)
- Amoebiasis
- Malaria

Toxic/drug induced

- Alcohol
- Drugs
- Poisons

Cryptogenic**Biliary tract disease/obstruction**

- Tumours
- Strictures
- Gallstones
- Biliary atresia

Miscellaneous

- Polycystic liver disease
- Congenital hepatic fibrosis
- Sarcoid
- Liver disease in pregnancy

Autoimmune

- Autoimmune hepatitis
- Primary biliary cirrhosis
- Primary sclerosing cholangitis

Vascular

- Budd–Chiari syndrome
- Portal vein thrombosis
- Veno-occlusive disease

Neoplastic

- Primary
 - Malignant
 - Benign
- Secondary

Helminthic

- Ascariasis
- Toxocariasis
- Clonorchis
- Schistosomiasis

autoimmune causes are infrequently encountered in the primary care setting, but in a tertiary centre make up a substantial proportion of patients. The standard ‘liver function tests’ (LFTs) are useful to determine that hepatitis is present, how severe it is, to document the progression of the disease and to assess the response to therapy. However, their role in identifying the aetiology of the inflammation is limited.

When the patient presents with an acute hepatitis, the liver function tests characteristically show the so-called ‘hepatic’ picture:

- a pronounced rise in plasma aspartate or alanine aminotransferase (AST, ALT) activity, often to >1000 U/L
- a modest (less than twice the upper limit of the reference range) increase in plasma alkaline phosphatase (ALP) activity
- bilirubinuria and, in more severe cases, hyperbilirubinaemia, which is clinically detectable as jaundice when the total plasma bilirubin rises to >40 – 50 $\mu\text{mol/L}$.

As a broad generalization, plasma activities of the aminotransferases reflect the severity of the disease. Certainly, patients who progress to fulminant hepatitis usually have exceptionally high plasma activities, with elevations of 20–40 times the upper reference limit. However, the relationship is not strictly quantitative, for lower activities (2–3 times the upper reference limit) can be found in

individuals with severe hepatic necroinflammatory activity on histological examination of liver biopsy material.

Differential diagnosis

There are two aspects to this part of the diagnostic process. The first is to distinguish between viral hepatitis and other, non-viral, causes of a hepatic illness and then, within each group, to identify the causative agent. Generally speaking, the standard LFTs are not helpful in separating viral from non-viral causes, and further serological testing, radiological imaging or histological examination will be required.

Plasma aminotransferase activities are not usually grossly raised in alcoholic hepatitis; in this condition, a ratio of AST:ALT of >2 is characteristic, while in hepatitis due to other agents the ratio is usually <2 . In fulminant hepatitis, aminotransferase activities may be close to the reference ranges by the time the patient reaches hospital, emphasizing that results of LFTs can change rapidly with time and depend on the particular clinical course that ensues. Generally speaking, acute infection with hepatitis B virus is more severe than that caused by hepatitis A, and biochemical abnormalities are more protracted. Raised plasma IgM concentration and atypical lymphocytes are more characteristic of type A, but the standard LFTs do not permit distinction between the different types of viral hepatitis.

Acute viral hepatitis

The specific diagnosis is established by serological tests for the hepatitis viruses (Table 14.1). The extent to which individual patients require biochemical and serological investigation depends on the clinical situation. For example, in children, a clinical diagnosis can be established with reasonable certainty if the clinical features are compatible with hepatitis A (infectious hepatitis), particularly if the physician knows of local occurrence. In adults, particularly where there has been no contact with infectious hepatitis, LFTs should be performed together with appropriate serology.

After a variable incubation period (the duration of which depends on the type of virus and the viral load), there is usually a rise in plasma aminotransferase activities, although some infants can be infected without any disturbance of LFTs. In many cases, particularly among young children, an elevated AST or ALT is the only indication of the hepatic process and the patient remains asymptomatic. When symptoms do develop, they are coincident with the maximal aminotransferase activities. Taking infection with the hepatitis A virus as an example, the patient begins to feel unwell and lethargic and may become anorexic, with transient diarrhoea being frequent. Pyrexia can develop and the liver is found to be enlarged and tender. The urine may become dark (owing to bilirubinuria), the stools pale, and in more severe cases, jaundice becomes evident a few days later. Symptoms tend to resolve with the onset of jaundice, which usually subsides over a few days. Following infection with this virus, the patient may, despite being asymptomatic, enter a cholestatic phase. This is heralded by rising plasma activities of ALP and γ -glutamyltransferase (γ GT) that can persist for many weeks and may cause considerable diagnostic confusion. Particularly in the older patient, it is an indication for ultrasound examination to rule out obstruction of the biliary tract. The jaundice usually begins to resolve before aminotransferase activities return

to the reference range, but it has been a frequent observation that bilirubin may disappear from the urine while the patient remains clinically jaundiced. The explanation for this situation lies in the development of 'bili-albumin' (see p. 239).

Outcome of acute viral hepatitis

There are three main possible outcomes of acute viral hepatitis, each with its own characteristic progression of clinical features and pattern of LFTs.

Complete resolution. As noted above, the plasma aminotransferase activities start to rise before the onset of jaundice and often before hepatomegaly is clinically detectable. Activities begin to fall coincidentally with the onset of jaundice and symptoms, and return to normal (AST before ALT) at the same time as, or shortly before, the plasma bilirubin concentration. This is the natural history in the great majority of patients and, unless jaundice develops, many patients will be unaware that they have had an acute hepatitis.

Progression to chronic liver disease. This mode of progression is limited to viral hepatitis types B and C, but a relapsing course for hepatitis A is well described and there are occasional reports of type A triggering autoimmune hepatitis in susceptible individuals. Persistence of symptoms, signs and/or abnormal liver tests (particularly aminotransferase activity in the range of 2–10 times the upper reference limit) for more than six months constitutes, by definition, chronic hepatitis (see below). The accompanying changes in viral serology are complex and will not be considered in the present discussion.

Progression to acute liver failure. Very rarely (in <1% of patients), any type of viral hepatitis may pursue a fulminant course. This course of events is characterized by grossly elevated plasma aminotransferase activities,

TABLE 14.1 Summary of differential diagnosis of acute viral hepatitis caused by true hepatotropic viruses

Nomenclature	Type of virus	Incubation period	Spread	Progression to chronic liver disease	Other features	Serological diagnosis
HAV	Picornavirus (RNA)	1–6 weeks	Faeco-oral	No	Raised IgM, atypical lymphocytes	IgM class anti-HAV
HBV	Hepadnavirus (DNA)	6 weeks–6 months	Parenteral	Yes		HBsAg and IgM class anti-HBc
HCV	Flavivirus (RNA)	2–26 weeks	Parenteral	Yes	Raised IgG and γ GT; polyphasic AST/ALT changes	Anti-HCV antibodies; detection of HCV-RNA by PCR
HDV (Delta)	RNA, HBV-dependent viroid	Uncertain	Mainly parenteral	Yes		IgM class anti-HDV
HEV	Calicivirus (RNA)	5–10 weeks	Faeco-oral	No	The major cause of enteric NANB	Anti-HEV (not yet routinely available)

HAV, hepatitis A virus; HBV, hepatitis B virus; HDV, hepatitis D virus; HEV, hepatitis E virus; NANB, non-A, non-B; PCR, polymerase chain reaction; for other abbreviations, refer to text.

an increasingly prolonged prothrombin time (PT) or international normalized ratio (INR) and the development of hepatic encephalopathy, which may occasionally develop before jaundice becomes a prominent feature. The prognosis is poor. The syndrome of acute liver failure (ALF), one cause of which is acute viral hepatitis, is described in more detail below.

ACUTE LIVER FAILURE

Acute liver failure implies the development of severe hepatic dysfunction within six months of the first onset of liver disease and in the absence of any pre-existing liver disease. Hepatic encephalopathy and a prolonged and persistent increase in PT are the characteristic features, and where these occur within eight weeks of the first symptom, the condition is known as fulminant hepatic failure (FHF). In those who survive, there are no long-term hepatic sequelae. The commonest cause in the UK is paracetamol overdose (Chapter 40) followed by viral hepatitis types B and E; ALF due to hepatitis C virus is extremely rare. Rarer causes include adverse drug reactions, other viruses, *Amanita* mushroom poisoning, Wilson disease, secondary hepatic malignancy, lymphoma and recreational drugs.

Laboratory features

Laboratory investigation is aimed at determining the cause of the ALF as well as assessing its consequences and complications (which are very similar irrespective of the aetiology). Measurement of the blood concentration of paracetamol is important, particularly in view of the beneficial effect of treatment by N-acetylcysteine infusion. The serology of viral causes is complicated because the hepatitis B surface antigen (HBsAg) may become undetectable before the patient reaches hospital. In general, with ALF due to causes other than paracetamol overdose, it is usually too late for specific treatment for the condition (e.g. D-penicillamine for Wilson disease) to be effective and management is by general supportive measures and, if feasible, liver transplantation.

The standard LFTs show a grossly hepatic picture at the time of onset of the encephalopathy. Jaundice is deep and progressive and plasma aminotransferase activities of several thousand U/L are usually found, although these may have fallen considerably by the time the patient is sent to a referral centre. A rare cause of FHF is massive metastatic tumour infiltration. This is characterized by hepatomegaly (in contrast to the reduced-size liver found with most other causes of FHF) and a much more cholestatic picture – such that the AST:ALP ratio is <4 rather than the converse, as in most other cases.

Coagulation defects are a consistent finding. Plasma concentrations of fibrinogen are decreased as are those of factors II, V, VII, IX and X. These are reflected in a prolonged PT or INR which, in the UK, is used as the main measure of severity of disease and is the most widely used parameter for following progress. Hypoglycaemia, owing to a combination of impaired gluconeogenesis and

glycogen breakdown and synthesis, is such a consistent finding that glucose infusion is a routine part of management, particularly during transfer to a referral centre. Where paracetamol is involved, it may interfere with blood glucose estimation, resulting in erroneously high values. Acute kidney injury is a frequent and ominous complication that occurs in at least 50% of patients with FHF, particularly with paracetamol poisoning. In about one-half of patients, the renal lesion is acute tubular necrosis, and in the remainder it is 'functional' renal injury (see p. 261).

Laboratory criteria for liver transplantation

Currently, with the best supportive care, 30–50% of patients with fulminant hepatic failure will survive; the figures are rather better for those with paracetamol overdose and rather poorer for those with viral hepatitis, particularly those classified as non-A–E, or drug reactions. This poor prognosis has led to the introduction of liver transplantation as a therapeutic option. It is clearly important to reserve this option for those with the worst prognosis and who are most unlikely to recover with supportive treatment. With this in mind, a number of laboratory-based criteria have been identified to ascertain the likelihood of death. The major adverse factors are an arterial hydrogen ion concentration >50 nmol/L (pH <7.3), and progression to grade III coma (see p. 259) with an INR of >10 and plasma creatinine concentration >300 µmol/L. Recently, there has been considerable interest in the measurement of plasma factor V concentration. If the ratio of factor V to factor VIII is very low, the chance of survival without transplantation is very small.

CHRONIC HEPATITIS

Chronic hepatitis is usually defined as a condition in which clinical or biochemical features of liver disease persist for more than six months. However, it has become clear that many patients can have the condition for much longer periods before it manifests itself in this way. Indeed, occasionally, such individuals may present with what at first appears to be an acute hepatitis that, on further investigation, proves to be an acute exacerbation of a previously asymptomatic chronic process (see below). On histological grounds, the condition was previously classified into two categories: chronic persistent hepatitis (CPH), defined as inflammation confined to the portal tracts with no necrosis of hepatocytes, and chronic active hepatitis (CAH) characterized by inflammatory cells spilling out into the hepatic parenchyma and damaging periportal hepatocytes in a hallmark pattern described as 'piecemeal necrosis'. Chronic persistent hepatitis was believed to be a benign condition of little clinical significance, with a good prognosis. Chronic active hepatitis, in contrast, was envisaged as a disorder with a more serious prognosis, which frequently progressed to cirrhosis, with the inflammatory activity often continuing even after cirrhosis developed (so-called 'active cirrhosis').

Although CPH and CAH were purely histological descriptions of the morphological changes seen in patients with chronic liver disease, they were embraced by hepatologists and gastroenterologists as clinical entities. However, during the 1970s and 1980s, it was increasingly recognized that the associated changes could be found in patients with several quite distinct causes of chronic hepatitis (Table 14.2). Somewhat surprisingly, rather than questioning the validity of CPH and CAH as distinct syndromes, this led to expansion of the list of diseases considered to be capable of causing them. By the 1990s, the previously accepted view that CPH was associated with a good prognosis was being challenged, because it became apparent that CPH and CAH are changes at either end of a spectrum of morphological changes seen in chronic hepatitis, and that transitions between them can and do occur during flares of disease activity and periods of remission, regardless of the aetiology of the process.

Other criteria for CPH and CAH were also found to be untenable. A requirement for markedly elevated plasma aminotransferase activities for diagnosis of CAH could not be upheld because it was recognized that these enzymes are poor correlates of histologically assessed activity in chronic liver disease, and that mild elevations of aminotransferases do not exclude severe disease. The temporal criterion for duration of disease of at least six months (to distinguish chronic from acute liver disease) also proved difficult to apply because it is often not possible to define the time of onset and, as noted above,

patients presenting with acute hepatitis who had clear evidence of chronic liver disease were being identified.

Current recommendations for defining and describing chronic hepatitis are that the terms CAH and CPH should be abandoned in favour of precise morphological descriptions, graded for necroinflammatory activity and staged for degree of fibrosis. While the term 'piecemeal necrosis' may be retained, the terms 'periportal hepatitis' or (preferably) 'interface hepatitis' should be used to describe the changes previously associated with CAH. Description of the changes previously associated with CPH should employ terms such as mild or moderate portal or periportal hepatitis (as appropriate) *without significant necrosis*. All of these terms should be qualified by aetiological designations (e.g. autoimmune hepatitis, chronic hepatitis B, C, D etc.), wherever possible and practicable.

Laboratory investigation plays a crucial role in the differential diagnosis and management of chronic hepatitis (see Table 14.2), although histological examination of liver biopsy material is usually required to assist in assessing severity and providing additional aetiological information. Typically, chronicity is defined as persistence of a hepatic pattern of abnormal LFTs over several months. The plasma aminotransferase activities are usually elevated to 2–10 times the upper reference limits. Plasma alkaline phosphatase is often normal or only slightly raised, although higher values may be seen in patients where cirrhosis has developed with associated distortion of the hepatic architecture. Plasma bilirubin concentration is often also normal or mildly elevated but, in severe cases, there can be a profound hyperbilirubinaemia with jaundice. The PT/INR and other markers of coagulation are often mildly abnormal, and the plasma albumin concentration may be at the lower end of the normal range, reflecting decreased hepatic synthetic function. Marked hypoalbuminaemia is, however, a late feature associated with advanced cirrhosis. Depending on the aetiology of the chronic hepatitis, plasma immunoglobulin concentrations may also be elevated, sometimes quite markedly (see below).

The above abnormalities may be a sequel to an acute hepatitis, in which case the aminotransferase activities may have risen to 10–20 times the upper reference limits before falling back to the somewhat lower values more characteristic of chronic progression. However, a clinically evident acute phase may not have been apparent especially if the patient did not develop jaundice. Indeed, chronic hepatitis is not infrequently diagnosed incidentally during routine health screening or investigation of some other condition. Notably, profound fatigue is a common feature of all forms of chronic hepatitis, and the latter should, therefore, be considered in the differential diagnosis of chronic fatigue syndromes.

TABLE 14.2 Differential diagnosis of chronic hepatitis

Type	Laboratory tests	Comments
Viral type B type C	HBsAg Anti-HCV; HCV-RNA by PCR	Has characteristic histological features
Alcoholic	Blood alcohol, γ GT, MCV, desialylated transferrin	Laboratory tests are supplementary to clinical history
Drugs	May develop autoantibodies	Oxyphenisatin, methyl dopa, isoniazid, dantrolene etc.
α_1 -Antitrypsin (AT) deficiency	Low α_1 -AT concentration; typical phenotype (PiZZ ^a)	Characteristic eosinophilic globules on histology
Wilson disease	Low caeruloplasmin, high tissue copper and urinary copper concentration	Low ALP activity; low AST/ALT activity for degree of inflammation
Autoimmune type 1	Anti-SMA and/or ANA	Markedly raised γ -globulin and IgG
type 2	Anti-LKM antibodies	

^aFor explanation, see α_1 -Antitrypsin deficiency on p. 268. ANA, antinuclear antibody; LKM, liver-kidney microsomal antibody; MCV, mean corpuscular volume; SMA, smooth muscle antibody.

Differential diagnosis of chronic hepatitis

Viral hepatitis types B and C

The specific diagnosis is based on virological tests. The detection in serum of the hepatitis B surface antigen (HBsAg) indicates that the patient is infected with the

hepatitis B virus (HBV). Clinical information and the results of LFTs can be useful in deciding whether the infection is acute, but this can be confirmed by the appearance of IgM antibodies against the HBV core antigen (HBc). The gradual disappearance of HBsAg, with sequential detection of IgG antibodies to HBc and antibodies to surface antigen (HBsAb) indicates clearance of the virus, an event that occurs in the majority of adults (90%) acutely infected with it. Other markers of viraemia are derived from different components of the virus, including hepatitis B 'e' antigen (HBeAg) and HBV DNA, and these disappear with viral clearance. The decline in HBe antigenaemia is accompanied by the appearance of antibodies to the 'e' antigen (HBeAb) and is known as seroconversion. In contrast to infection acquired in adulthood, perinatal infection becomes chronic in 95% of patients, with the virus persisting into adult life. Chronic infection is associated with a number of outcomes. Many individuals are asymptomatic, with the virus detected incidentally and a serological profile in which HBsAg and HBeAb are detected and HBeAg is negative. Although such people frequently have normal liver function tests, and have long been termed 'healthy carriers' and been considered to be of low infectivity, it has become clear that a significant proportion have marked histological damage on liver biopsy and detectable circulating HBV DNA. Others have clear evidence of cirrhosis and, in such individuals, there is a 5% annual risk of developing hepatocellular carcinoma.

Hepatitis C virus (HCV) infections are identified initially by detection of antibodies (anti-HCV) to the virus and chronic infections confirmed by seropositivity for the viral RNA (HCV-RNA).

Antiviral therapy for chronic hepatitis B and C, using interferons, with or without other antiviral agents to inhibit viral replication, is now well established, and comprehensive guidelines for treatment exist. Thus, tests for genomic material of both viruses (HBV-DNA and HCV-RNA, respectively) are usually performed, initially to assess the viral load before treatment and subsequently to monitor the patient's response to treatment. Plasma aminotransferase activities are also usually monitored as an adjunct to the above, since they are cheap to perform on a frequent basis and a return to normal values suggests at least some response to treatment. In the case of HBV, successful treatment is heralded by a change from HBeAg to HBeAb positivity and this event is accompanied by an acute hepatic illness, the so-called 'hepatic flare' (Fig. 14.1). Hypothyroidism is a side effect of interferon therapy, particularly in patients with chronic hepatitis C, and thyroid function tests should be monitored before treatment and at 3-monthly intervals.

Procedures for the handling of potentially dangerous and infectious samples are not the concern of this chapter, but it is worth noting that as many as 1–2% of inner city populations may be carriers of either HBV or HCV. The majority of these will have no symptoms of liver disease and their blood samples may be referred to the laboratory for causes unrelated to liver disease. Caution should therefore be exercised in the handling of all blood products in view of the possibility of occult hepatitis viral infections.

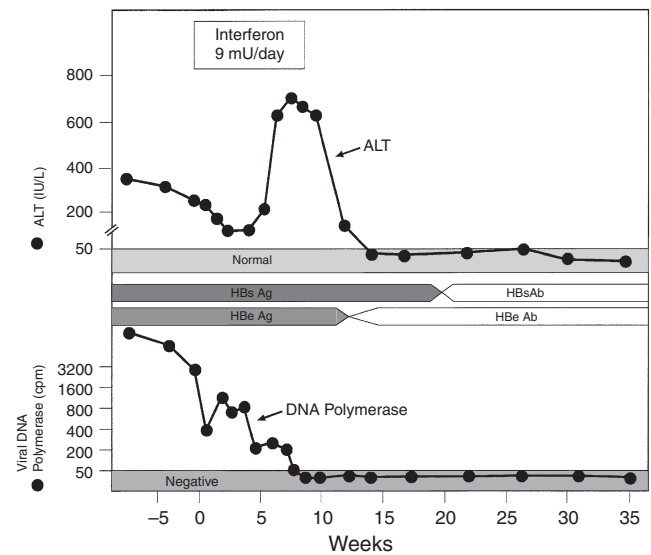


FIGURE 14.1 ■ Changes in ALT activity in a patient with chronic hepatitis B virus infection successfully treated with interferon. Note the hepatitis response heralding clearance of HBsAg and appearance of the antibody. Clearance of the virus is also indicated by a decrease in viral DNA polymerase activity.

Alcohol

In perhaps 10% of patients with alcoholic liver disease, the histological features of interface hepatitis, similar to those in other forms of moderate/severe chronic hepatitis can be observed. There is evidence that some such cases may be related to coexistent hepatitis C virus infection or other underlying causes. The problem that this presents for differential diagnosis is discussed in detail under the heading of alcoholic liver disease.

Wilson disease

Chronic hepatitis is a rare presentation of this rare disease (see p. 267), but is of considerable importance because treatment can be instituted and further liver damage avoided. Generally, the plasma activities of aminotransferases are low in comparison with the extent of histological activity. The diagnostic work-up is considered in detail below, but the initial screening procedures include measurement of plasma copper and caeruloplasmin concentrations and ocular slit-lamp examination for Kayser–Fleischer rings. The diagnosis must be considered in all patients with chronic hepatitis, particularly those in the younger age groups.

α_1 -Antitrypsin deficiency

This is another rare cause of chronic hepatitis (see p. 266). Plasma α_1 -antitrypsin concentrations are usually low but, occasionally, because of active inflammation, they may be within the reference range. It is therefore essential that the α_1 -antitrypsin phenotype is determined, particularly in younger patients.

Autoimmune hepatitis (AIH)

This is a rare cause of chronic hepatitis. However, it is one of the few causes of chronic liver disease in which

drug treatment is highly effective in the great majority of patients, and accurate diagnosis is essential. This is established partly by careful exclusion of the other causes listed above and partly by the finding of a suggestive pattern of biochemical, immunological and histological abnormalities. The biochemical liver tests show a hepatitic pattern (see above), but plasma aminotransferase activities are often only moderately elevated and bilirubin concentrations are frequently normal. Typically, there is a marked hypergammaglobulinaemia with selective elevation of the IgG concentration (which can be as high as 100 g/L). Approximately 80% of AIH patients will also have significant titres (>1:40) of organ non-specific autoantibodies (smooth muscle antibody (SMA), antinuclear antibody (ANA) or anti-liver-kidney microsomal (LKM) antibodies). Standard treatment involves immunosuppressive therapy with prednisolone and the steroid-sparing agent azathioprine.

Monitoring response to therapy. Plasma aminotransferase activities do not correlate well with histologically assessed disease activity in AIH. However, such tests, together with plasma IgG concentrations, are cheap and can be repeated more easily than liver biopsy, so are used routinely to monitor response to immunosuppressive therapy. Some 80–95% of AIH patients respond to the standard therapy of prednisolone and azathioprine, with aminotransferase activities and IgG concentrations falling to at least 50% below their initial values within three months. Dosages are titrated against the aminotransferase and IgG values, being progressively reduced to maintenance doses as these parameters fall towards the normal range that, in most patients, is achieved within one year of starting treatment. However, this 'biochemical remission' does not represent a complete response. It usually requires at least a further year of treatment before complete histological remission is achieved. Current recommendations are that treatment should be continued for at least two years (preferably four), with regular monitoring of aminotransferase activities and IgG concentrations; before any attempt is made to withdraw treatment, histological remission should be confirmed. Withdrawal of all immunosuppression is rarely achieved, possibly because it is not routinely attempted, but gradual reduction in the dose of prednisolone to very low levels, or complete withdrawal, is feasible so that immunosuppression can be maintained with azathioprine alone. If complete withdrawal of immunosuppressive treatment is achieved, the patient should be monitored indefinitely, because even a complete histological response does not represent a true 'cure' and there is a lifelong risk of relapse of the disease. The latter is heralded by a rise in aminotransferase activities and IgG concentrations. Importantly, interpretation of biochemical test results should not focus solely on the upper reference limits for the various parameters but should take account of what may be normal for the individual patient. By definition, half of the normal population will have values below the geometric means of the reference ranges and, for such individuals, a rise in a parameter to (or just above) the upper reference limit may represent a doubling of their normal value.

PRIMARY BILIARY CIRRHOSIS (PBC)

This is a chronic cholestatic condition of unknown cause in which there is destruction of septal and interlobular bile ducts. Females are affected far more frequently than males. Most patients present with pruritus, jaundice or non-specific symptoms including tiredness and hepatic pain. However, an increasing number are detected incidentally when abnormal liver tests, particularly an increased activity of alkaline phosphatase, are detected during opportunistic screening. Although PBC is considered to be an autoimmune liver disease, and twin and family studies suggest that there is a significant genetic component, it is clearly a complex disease and little progress has been made in defining the genes responsible.

Biochemical liver function tests reveal a characteristic cholestatic picture with, as the disease progresses, increasing plasma alkaline phosphatase (ALP) (of biliary origin), increasing bilirubin concentration and falling albumin. The plasma concentrations of both pentameric and monomeric IgM are frequently raised (*cf.* AIH). The most specific serological tests are those for anti-mitochondrial antibodies (AMA), which are detectable in about 95% of patients. Antimitochondrial antibodies recognize several distinct autoantigens, but the most characteristic for PBC are the M2 antibodies that react with epitopes on the E2 components of the pyruvate dehydrogenase complex (PDC-E2), the branched chain oxoacid dehydrogenase (BCOAH-E2) and the oxoglutarate dehydrogenase complexes. Laboratory tests, particularly the plasma bilirubin concentration, are used in assessing prognosis (see p. 248).

PRIMARY SCLEROSING CHOLANGITIS (PSC)

This is a progressive disease characterized by diffuse inflammation and fibrosis of the extrahepatic and/or intrahepatic biliary system, which leads to obliteration of the intrahepatic ducts and, eventually, biliary cirrhosis (see below). It is diagnosed on the basis of characteristic cholangiographic appearances together with compatible clinical, biochemical and histological features, and exclusion of several other conditions known to cause secondary sclerosing cholangitis, such as biliary calculi and previous biliary tract surgery. Men are affected more often than women and, again, although considered an autoimmune disease, the disease is genetically complex and remains poorly characterized.

With the advent of endoscopic retrograde pancreatography (ERCP), PSC began to be diagnosed earlier in its natural history, particularly following the initial finding of persistently abnormal LFTs in patients with inflammatory bowel disease, which coexists with PSC in more than 50% of cases. In the early stages, there is usually a minor elevation of plasma aminotransferase activities and moderate elevation of ALP, which progresses, over several years, to a very severe cholestatic condition with deep jaundice. By this stage, plasma ALP activity often reaches more than ten times the upper reference limit. Death from liver failure ultimately results unless a liver

transplant can be performed. However, the disease has been known to recur in the grafted liver. Rapid clinical deterioration with a progressive cholestatic picture may indicate the development of cholangiocarcinoma, a complication with a lifetime risk of 10–15% in this condition. CA19-9 concentrations may be elevated when a cholangiocarcinoma has developed but this is not a specific finding as CA19-9 is also increased in other cholestatic liver diseases. Copper retention appears to be a general feature of chronic cholestatic conditions: in PSC most patients will have raised hepatic copper concentrations (of a similar order to that seen in Wilson disease and primary biliary cirrhosis) and urinary copper concentrations are typically raised. However, in contrast to Wilson disease, the plasma caeruloplasmin and copper concentrations will remain within the reference range (see p. 267).

ALCOHOLIC LIVER DISEASE

Excessive alcohol (ethanol) consumption is by far the commonest cause of liver disease in the Western world, although it often exists as part of a wider spectrum of social, psychological and pathological effects of alcohol-related damage. Most current evidence suggests that alcohol itself is the primary cause of the liver damage, although associated malnutrition may be a contributory factor. Considering that there is an enormous variation in the amount of alcohol drunk by different individuals, the time course over which it is consumed and individual susceptibility to tissue injury, it is not surprising that the adverse effects on the liver vary widely.

Ethanol metabolism

Alcohol is absorbed from the stomach and small bowel. Absorption is most efficient when there is no other food (particularly carbohydrate) in the gut and when its concentration in the ingested fluid is of the order of 20%.

The liver is exposed to the highest concentration of alcohol in blood, owing to the blood supply by the hepatic portal vein, and is responsible for more than 95% of its metabolism. The concentrations attained after ingestion of a standard amount of ethanol depend, among other things, on sex, weight, previous exposure to alcohol, the type of alcoholic beverage and rate of gastric emptying. Ethanol is oxidized to acetaldehyde mainly by the cytosolic enzyme alcohol dehydrogenase, but also, particularly at high concentrations, by the cytochrome P450 system. Acetaldehyde is particularly toxic. It is metabolized by aldehyde dehydrogenase in the mitochondria to acetate, which is, in turn, oxidized by peripheral tissues to carbon dioxide and water. Both enzyme systems can be induced by alcohol. The intoxicating and metabolic effects are mediated directly by ethanol but it is likely that acetaldehyde is an important factor in causing tissue damage through the formation of adducts that in turn lead to functional impairment of hepatic proteins. The metabolism of ethanol also predisposes to various metabolic problems including hypoglycaemia and non-respiratory acidosis (see Chapter 40).

Liver pathology in alcoholic liver disease

Three patterns of histological change in liver tissue have been described in association with ethanol ingestion, although there is considerable overlap and all three features can be present on occasions. Fatty liver (steatosis) appears to be a response to excessive alcohol consumption in all individuals. It is considered to be a reversible feature, resolving with abstinence, and does not progress to chronic liver disease provided abstinence is maintained. In alcoholic hepatitis, the appearances are of steatosis but also sometimes with features of interface hepatitis. This is a serious condition with a marked propensity to progress to fibrosis and cirrhosis – the third and final stage of chronic alcoholic liver disease. Additional morphological changes may be seen in patients who have other coexisting disorders. For example, people with hereditary haemochromatosis or chronic hepatitis C may be at increased risk of developing liver disease from alcohol, and porphyria cutanea tarda is often associated with excessive alcohol consumption.

Biochemical abnormalities

Alcoholic steatosis. Biochemical changes are minimal and consist of subclinical hyperbilirubinaemia and a very mild elevation of plasma aminotransferase activities. Very occasionally, a cholestatic syndrome develops, but associated alcoholic pancreatitis leading to biliary obstruction should be considered if jaundice is marked. Episodes of delirium tremens and alcoholic myopathy may give markedly raised plasma aminotransferase activities. Plasma γ GT activity is elevated in the majority of patients (although it is not a specific finding), but usually reflects enzyme induction rather than hepatic injury.

Alcoholic hepatitis. There is a wide spectrum of severity associated with this histological diagnosis, ranging from the classic syndrome of deep prolonged jaundice, hepatic failure, fever and leukocytosis through to a complete absence of symptoms and physical signs. Alcoholic hepatitis generally occurs after heavy bouts of drinking and when subjects have been drinking heavily for several years. The laboratory tests reveal anaemia, usually with a leukocytosis and, consistently raised aminotransferase activities. However, the AST values are seldom above ten times the upper reference limit and ALT values are usually lower. This results in an AST:ALT ratio of >2 , and a value below this in subjects with hepatitis suggests that alcohol is not a major aetiological factor.

Alcoholism and haemochromatosis. Iron overload is common in alcoholic liver disease, occurring in perhaps 50% of patients (see p. 265). Occasionally, this may progress to a similar degree to that seen in hereditary haemochromatosis, and the patient may have similar abnormalities in biochemical parameters of iron status and endocrine function. More commonly, it appears that excessive alcohol consumption leads to the unmasking of haemochromatosis in patients with this genetic disease.

Porphyria cutanea tarda. As with haemochromatosis, this genetically determined condition, comprising bullous skin lesions on exposure to sunlight, liver disease and iron overload, is often brought to medical attention by excessive alcohol consumption. The underlying biochemical defect is a deficiency of uroporphyrinogen decarboxylase (see Chapter 28).

Use of laboratory tests in clinical practice

Biochemical tests of alcohol intake and its misuse fall into three categories:

- measurement of alcohol or its metabolites, e.g. ethanol, ethyl glucuronide, ethyl sulphate, 5-hydroxytryptophol
- assessment of the effect of alcohol on protein metabolism, i.e. carbohydrate deficient transferrin
- assessment of the extent of liver damage associated with alcohol misuse by measurement of plasma enzymes.

Alcohol and metabolites

Saturation of the normal pathways of alcohol metabolism results in the induction of minor pathways with the formation of metabolites with high specificity for excess alcohol intake: glucuronidation leading to ethyl glucuronide (EtG) formation; sulphation to ethyl sulphate (EtS) formation, and incorporation into 5-hydroxytryptamine to form 5-hydroxytryptophol. Ethanol in blood or urine remains detectable only for 24–36h, even after ingestion of significant amounts, so false negative results may be obtained in subjects misusing alcohol if they can abstain for 1–2 days prior to testing. The urinary 5-hydroxytryptophol:5-hydroxyindole acetic acid ratio has close to 100% specificity for excess alcohol consumption, but mass spectrometry is required for its measurement and the test has not been adopted routinely. Recently, interest has grown in the use of EtG and EtS measurements as these metabolites have high specificity for alcohol misuse and can be detected in urine up to 72–90h after significant intake. A commercial immunoassay exists for EtG, but the combined measurement of EtG and EtS requires mass spectrometry. Many workers favour the latter approach because of reports of both false positive and negative EtG results in the presence of a urinary infection with bacteria capable of either removing the glucuronide or glucuronidating ethanol *in vitro*. Ethyl glucuronide and EtS do, however, offer the potential for detection of ‘binge’ drinkers that has been notoriously difficult to achieve in the past.

Effects of alcohol on protein metabolism

Carbohydrate deficient transferrin (CDT) is defined as the asialo- and disialo- forms of transferrin that usually account for <1.6% of the total transferrin found in plasma. The formation of CDT is directly proportional to alcohol intake: ethanol (or its metabolites) appears to inhibit the enzymes responsible for addition of the carbohydrate side-chains and induce sialidase that removes the terminal sialic acid residues from the side-chains.

The percentage CDT is independent of the extent of any alcohol-related liver damage except in patients with severe cholestatic liver disease, where reduced clearance may cause a falsely high result. Hepatic failure can cause analytical difficulties in CDT measurement because of low total transferrin concentrations from reduced synthesis. As the plasma half-life of CDT is 10–14 days, a raised percentage of CDT is strongly suggestive of chronic excessive alcohol consumption with sensitivity and specificity both approaching 85%. Measurement of serum immunoglobulin concentrations may also be helpful because alcoholic liver disease is often associated with a selective increase in IgA (*cf.* AIH and PBC).

Plasma enzymes

Alcoholic steatosis can cause an increase in the aminotransferases or γ -GT that is reversible with abstinence, although the degree of increase correlates poorly with the extent of liver damage assessed histologically. Progression to cirrhosis is usually accompanied by abnormal plasma enzyme activities often suggesting intrahepatic cholestasis, although in some patients with established cirrhosis who abstain from further alcohol intake, these may revert to normal. For a more detailed discussion, including the role of the measurement of mitochondrial AST, see Chapter 13.

Non-alcoholic fatty liver disease (NAFLD)

This comprises a spectrum of conditions ranging from simple hepatic steatosis (excessive accumulation of fat in hepatocytes) to end-stage chronic liver disease: these conditions are clearly associated with the epidemic of obesity and diabetes recognized in Western populations. Previously thought to be a rare and relatively benign condition, developing mainly in middle-aged females in association with obesity, insulin resistance and hypertriglyceridaemia, NAFLD is emerging as one of the commonest causes of chronic liver disease. Steatosis is an adaptive response of the liver to insulin resistance. When this is accompanied by endogenous insults such as oxidative and free radical damage, mitochondrial dysfunction and endotoxaemia, themselves resulting in part from excessive amounts of fat in liver cells, an inflammatory condition termed non-alcoholic steatohepatitis (NASH) is induced in susceptible individuals. Typical histological features include micro- and macrovesicular steatosis, mild to moderate portal and lobular inflammation (often with small clusters of polymorphs), liver cell ballooning and perisinusoidal fibrosis. Mallory’s hyaline (typically associated with alcoholic liver disease) may also be present. Non-alcoholic steatohepatitis can induce fibrosis and, in some individuals, this process will result in cirrhosis with all the attendant complications including hepatocellular carcinoma. Unfortunately, it is not possible at present to predict which individuals will experience a progressive course. Liver function tests are mildly or moderately abnormal in most patients at presentation, but longitudinal studies indicate that progression of fibrosis can occur despite return to normal of aminotransferase activities in response to weight loss.

THE CONCEPT OF CIRRHOSIS

Cirrhosis is not a clinical diagnosis, but a pathological description of the liver in which there is:

- diffuse hepatic fibrosis
- nodular regeneration
- a disturbance of the normal hepatic architecture, that is, a distortion of the normal relationship of the portal tracts to the central veins (see Chapter 13).

It is the end result of chronic liver diseases that are usually associated with recurrent episodes of necrosis, cell death and attempts by the liver to regenerate. Liver function is disturbed, not only because of the cirrhosis per se, which has little effect on standard LFTs, but because of continuing immunologically mediated liver cell damage, the capillarization of the sinusoids due to fibrosis (see Chapter 13) and, on occasions, because of the effects of the initiating agent such as alcohol. Efforts to make a diagnosis of cirrhosis on any grounds other than histological are unreliable.

A great part of clinical hepatology is taken up with the diagnosis and management of the complications of cirrhosis: hepatic encephalopathy, ascites and the hepatorenal syndrome, infections, primary hepatic malignancy and endocrine dysfunction.

Hepatic encephalopathy

This is seen in patients with advanced cirrhosis, and a precipitating factor such as administration of a sedative, a large protein intake (including blood from variceal haemorrhage), infection or electrolyte imbalance can usually be identified. It is particularly frequent in subjects who have undergone a portocaval shunt for treatment of portal hypertension. As with the encephalopathy associated with acute liver failure, the severity of the condition is graded from I–IV:

- I – alert but with asterix ('hepatic flap'), inversion of sleep rhythm
- II – confused, disorientated, slow mentation, inappropriate behaviour
- III – restless, aggressive, sleepy, uncommunicative
- IV – coma.

The diagnosis is essentially a clinical one and, although characteristic electroencephalographic appearances are described, these are not used diagnostically. Similarly, the blood ammonia concentration is usually raised but correlates only poorly with the degree of encephalopathy.

Vascular disturbances in cirrhosis

Cirrhosis is characterized by the development of systemic vasodilatation, which results in a decrease in effective arterial blood volume, low systemic vascular resistance and a hyperdynamic circulation with high cardiac output. The mechanism underlying these changes is not known with certainty, but may involve increased vascular synthesis of nitric oxide because of increased activity of inducible nitric oxide synthase, as well as increased production of prostacyclin and other

vasodilators such as glucagon, substance P and calcitonin gene-related peptide. In response to vasodilatation, several homeostatic systems are triggered, including an increase in renal sympathetic activity, activation of the renin–angiotensin–aldosterone system and increased production of antidiuretic hormone (ADH). In some patients, vascular resistance and effective arterial blood volume are restored, but in others, particularly when cirrhosis is complicated by severe portal hypertension, water retention and vasoconstriction are insufficient to restore circulating blood volume and these subjects present with hypotension and a hyperdynamic circulation, together with high plasma concentrations of renin, noradrenaline (norepinephrine) and ADH. These hormones have a profound effect on renal vasculature, causing progressive vasoconstriction and a reduction in plasma flow and glomerular filtration, resulting in marked salt and water retention. These changes have a central role in the development of ascites.

Ascites

Ascites, the excessive accumulation of extracellular fluid in the peritoneal cavity (over 30L in some patients), is usually a late complication of cirrhosis. Two key factors underlie its development, namely portal hypertension operating at the level of the sinusoids, and sodium and water retention. Portal hypertension occurs as cirrhosis develops because of progressive collagen deposition within the space of Disse and nodule formation, resulting in disruption of the vascular architecture of the liver. Increased resistance to portal flow and increased hydrostatic pressure within the sinusoids favour transudation of tissue fluid into the peritoneal cavity. Portal hypertension is a critical permissive factor in the development of ascites. The second key factor in the development of ascites is, as described above, salt and water retention with enhanced sodium reabsorption occurring in both the proximal and distal tubules of the kidneys. In some patients with ascites, normal concentrations of aldosterone are found, leading to the suggestion that enhanced renal sensitivity to the hormone underlies the increased distal tubular sodium reabsorption. The role of hypoalbuminaemia in the development of ascites in patients with liver disease is controversial. Some consider this to have a permissive role in localizing the fluid collection to the peritoneal space, while others consider that plasma albumin concentration has little influence on ascites formation. Instead, they suggest that the oncotic pressure gradient across the sinusoids is low (because albumin crosses the fenestrations) so ascites develops only when trans-sinusoidal filtration exceeds the drainage capacity of the lymphatic system, as occurs in patients with portal hypertension.

Ascites usually develops in patients with advanced, decompensated disease, but the tendency to retain salt is present before ascites actually develops and a sudden increase in dietary salt intake may lead to the generation of ascites that clears spontaneously when the salt intake is restricted. Excess water retention leading to hyponatraemia is frequent and is attributable to the mechanisms described above. There are, however, several liver and

non-liver diseases other than cirrhosis (Box 14.2) that can lead to the formation of ascites, and the establishment of a precise diagnosis is important for its correct management.

If the standard LFTs are abnormal in a patient with ascites, then it is likely that one of the liver diseases listed in Box 14.2 is responsible and it is often already known that a patient who develops ascites has cirrhosis. However, patients with liver disease can develop ascites for reasons other than cirrhosis. For example, patients with alcoholic cirrhosis may develop tuberculous ascites, pancreatic ascites or 'malignant' ascites if hepatocellular carcinoma supervenes. Cardiac causes can usually be diagnosed confidently on clinical grounds, but the other causes may be difficult to distinguish and laboratory investigations have an important role to play. Ascitic fluid should be aspirated using an 18-gauge needle and syringe (a so-called 'diagnostic tap') for analysis. The investigations and their interpretation are shown in Table 14.3. It should be recognized that sodium retention is not a diagnostic feature of cirrhotic ascites: equally intense sodium retention can occur in malignant ascites.

The figures for ascitic protein given in Table 14.3 are given only as guidelines. Exceptions are frequent and clinical utility is limited, but as the changes become more extreme, so their diagnostic specificity increases. Thus, a protein concentration <10 g/L is not infrequent in

uncomplicated cirrhosis, but virtually rules out malignant disease. Conversely, protein concentrations >35 g/L are the rule in malignant ascites and are unusual in uncomplicated cirrhosis.

Several tests have been proposed to increase the specificity of total ascitic protein measurement in differentiating between cirrhotic and malignant ascites. These include calculation of the plasma to ascites albumin gradient and measurement of ascitic lactate dehydrogenase activity or ascitic cholesterol concentration (Fig. 14.2). Adenosine deaminase activity in ascites of >60 U/L is particularly sensitive and specific for tuberculous ascites.

Monitoring treatment of ascites

Ascites due to cirrhosis is usually managed by a combination of dietary salt restriction and diuretic therapy, though paracentesis with albumin infusion is now being used more widely. The aim should be to achieve a net fluid loss of 500 mL/24h until the ascites clears. Plasma urea, creatinine, sodium and potassium concentrations should be checked daily at the start of treatment. A rising plasma urea or creatinine and falling plasma sodium concentration (<130 mmol/L) indicate impending acute kidney injury and should prompt a reduction of diuretic treatment. Diuretics should not be used when plasma sodium is <125 mmol/L. Plasma creatinine concentration is the more useful test, as impaired urea synthesis in patients with cirrhosis makes the plasma urea a less sensitive indicator of renal function. Plasma potassium needs careful monitoring if furosemide is used because of the risk of hypokalaemia, which may precipitate encephalopathy. So, too, can hyperkalaemia, which may occur with spironolactone if renal impairment supervenes.

Acute kidney injury

The onset of acute kidney injury (AKI) is indicated by rising plasma concentrations of creatinine and urea and usually, but not always, a urinary output falling to <300 mL/24h. A particularly dangerous complication, as already mentioned, is hyperkalaemia. Such developments are likely to precipitate encephalopathy in patients with liver disease. The major clinical problem in such patients is an idiopathic form of AKI that is widely known as the 'hepatorenal syndrome' (see below) and usually associated with advanced liver disease, ascites and encephalopathy. This should not be confused with

BOX 14.2 Hepatic and non-hepatic causes of ascites

Primary hepatic disease

- Cirrhosis (of any type)
- Budd–Chiari syndrome (thrombosis of the hepatic veins)
- Acute and subacute liver failure
- Portal vein thrombosis – rarely, and late in natural history

Non-hepatic disease

- Abdominal malignancy
- Cardiac failure
- Constrictive pericarditis
- Nephrotic syndrome
- Peritonitis (especially tuberculous)
- Pancreatic disease
- Protein–calorie malnutrition
- Hypothyroidism (rarely)

TABLE 14.3 Differential diagnosis of ascites based on inspection and laboratory tests

	Direct inspection	Protein content (g/L)	Cells	Bacteria
Uncomplicated cirrhosis	Clear	<25	<500/mm ³ (mononuclear cells)	None
Intra-abdominal malignancy	Clear	>25	Malignant cells (and red blood cells)	None
Tuberculous peritonitis	Turbid	>25 ^b	White blood cells (mainly lymphocytes)	Acid-fast bacilli
Abdominal lymphoma	Chylous ^a	See notes below ^c	None	None

^aChylous ascites probably represents escape of splanchnic lymph into ascites.

^bConcentrations may be lower when cirrhotic ascites is complicated by tuberculosis.

^cThe characteristic laboratory feature is a high concentration of triglycerides.

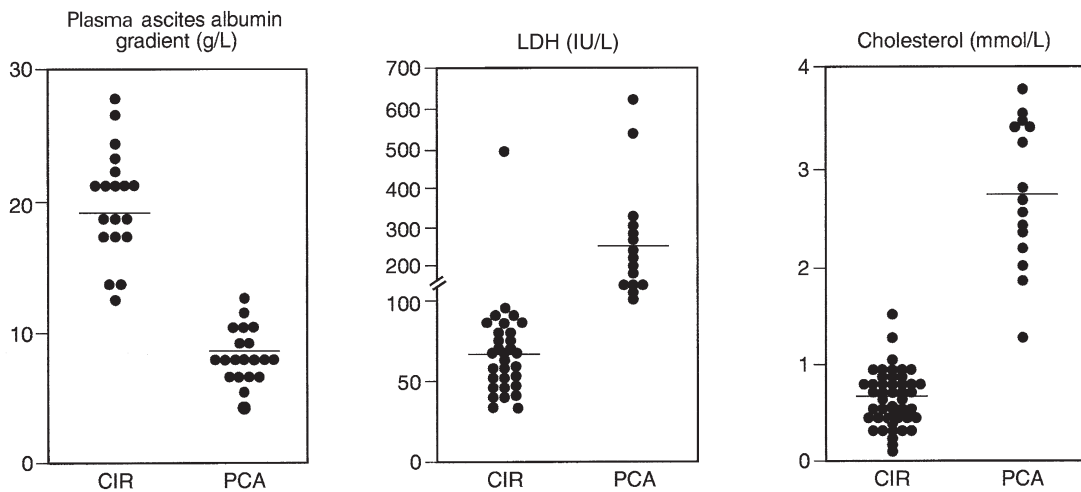


FIGURE 14.2 ■ Plasma ascites albumin gradient and ascitic fluid concentration of lactate dehydrogenase (LDH) and cholesterol in patients with cirrhosis (CIR) and peritoneal carcinomatosis (PCA). Reproduced with permission from Arroyo V et al. Ascites, renal failure and electrolyte disorders in cirrhosis. Pathogenesis, diagnosis, and treatment. In: McIntyre N et al. (eds). Oxford Textbook of Clinical Hepatology. Oxford: Oxford University Press; 1991.

several other situations in which renal disease and liver disease coexist: polycystic disease, infections such as leptospirosis, circulatory failure and the immune complex glomerulonephritis associated with hepatitis B virus infection or the presence of cryoglobulins in hepatitis C infection. Acute kidney injury may also develop when cirrhotic patients are fluid depleted, and following surgery to relieve obstructive jaundice. The latter syndrome is caused by acute tubular necrosis, but the mechanism is unknown.

The hepatorenal syndrome (HRS)

This condition is also known as functional renal failure (FRF). The characteristic feature is that the kidneys are normal when examined histologically, that is, there is disturbance of function but not of structure; specifically, the usual appearances of acute tubular necrosis are not present. Furthermore, the impairment is progressive and not reversed by fluid repletion, suggesting that this is not simply a pre-renal disorder. The precise pathogenesis is unknown, but underlying factors include those detailed in the earlier section on vascular abnormalities in cirrhosis, which lead to intense renal vasoconstriction resulting in decreased renal blood flow and a reduced glomerular filtration rate. Functional renal failure occurs classically in patients with advanced chronic liver disease, but may also occur in severe acute liver disease.

In patients known to have advanced cirrhosis, the condition can be diagnosed by simple laboratory tests and clinical observation. Reduced urine flow and rising plasma concentrations of urea and creatinine are common to all forms of AKI, but the characteristic feature of FRF, which distinguishes it from acute tubular necrosis, is the dramatic degree of sodium retention (Table 14.4). The urinary sodium concentration is usually <12 mmol/L and often <5 mmol/L. The urine:plasma osmolality is usually >1, and this contrasts with patients with acute

TABLE 14.4 Differential diagnosis of renal impairment in cirrhosis

	Prerenal disease	Functional renal failure	Acute tubular necrosis
Urinary sodium concentration (mmol/L)	<12	<12	>12
Urine:plasma osmolality	>1.15	1.1–1.15	<1.1
Response to volume expansion	Yes	No	No

It must be emphasized that although these tests are typical, exceptions do occur, particularly in patients who have been recently treated with diuretics.

tubular necrosis, in whom the urinary sodium concentration is invariably >12 mmol/L and often >20 mmol/L, and the urine/plasma osmolality is about 1. Infusion of terlipressin, a non-selective agonist of V_1 vasopressin receptors, together with albumin, reverses FRF in ~75% of patients.

Sex hormones and their binding proteins

Men with cirrhosis are frequently impotent, infertile and feminized (i.e. they may have gynecomastia, a female distribution of body hair and testicular atrophy). Workers in the USA, where alcohol is responsible for a very high percentage of all chronic liver disease, have tended to attribute these symptoms to the alcohol rather than the chronic liver disease. In the UK, where at least 50% of patients with cirrhosis have non-alcohol related disease, sexual dysfunction is still frequently seen and it seems most likely that alcohol may act both on its own and in concert with chronic liver disease to cause this disorder. The biochemical and endocrine changes

that accompany chronic liver disease are now well recognized but they correlate poorly with the clinical symptoms.

This section describes the changes in endocrine tests that are found in patients with liver disease of various causes and severity, but it is important to note that the changes described are not necessarily the cause of symptoms. The discussion below refers mainly to men; the effect of liver disease on sexual function in women has been less well studied.

Physiology and biochemistry

This topic is discussed in detail in Chapter 23 but a summary is presented here for convenience. The major circulating androgen is testosterone secreted by the testicular Leydig cells (0.17–0.35 mmol/24 h) under the control of luteinizing hormone (LH). The androgenic effect of androstenedione, and other adrenal androgens such as dehydroepiandrosterone, is probably attributable to their peripheral conversion to testosterone. Testosterone can be metabolized to its more active metabolite dihydrotestosterone (DHT) by 5α -reduction, or to oestrogens (by aromatization) or can be degraded in the liver, which is responsible for clearing about 50% of testosterone during the first pass in males. Oestrogens also circulate in men, the most potent being oestradiol (strictly 17β -oestradiol, but often abbreviated to E₂), which is derived from aromatization of testosterone. Testosterone and 17β -oestradiol circulate bound to albumin (low affinity, unsaturable binding) and sex hormone-binding globulin (SHBG, high affinity, saturable binding). Only a small amount of each hormone exists in the unbound state and this is presumed to be the biologically active fraction. Spermatogenesis is under the control of follicle stimulating hormone (FSH), although LH is also required, acting indirectly on the seminiferous tubules by stimulating testosterone secretion by the adjacent Leydig cells.

Changes in men with cirrhosis

The plasma total testosterone concentration is at the lower limit of the reference range in most men with well-compensated liver disease and falls as the disease progresses. The oestradiol concentration, on the other hand, is usually at the upper limit of the reference range and rises as the disease progresses, and the ratio of oestradiol to testosterone is thus elevated. Because SHBG concentrations are invariably also raised and the affinity of SHBG for testosterone is considerably greater than that for oestradiol, the ratio of free oestrogen to free testosterone is even greater. The low testosterone concentrations are attributable to reduced testicular production and occur despite a decrease in metabolic clearance rate caused by the high SHBG concentrations. The elevated oestrogen concentrations are not due to impaired hepatic clearance as originally thought; increased production, probably due to enhanced peripheral aromatization, appears to be the most likely mechanism.

In the presence of low testosterone concentrations, a compensatory increase in pituitary secretion of

TABLE 14.5 Changes in plasma sex hormone concentrations in men with chronic liver disease in relation to its severity

	Well compensated cirrhosis	Poorly compensated cirrhosis
Testosterone (T)	Low/normal	Low
17β -oestradiol (E ₂)	High/normal	High
SHBG	High	High
Total E ₂ /T ratio	High	Very high
LH	Normal	Low

luteinizing hormone (LH) would be expected, but this does not always occur, implying that a primary testicular defect is often complicated by hypothalamic–pituitary dysfunction. As the disease progresses, LH concentrations tend to fall. Thus, the characteristic abnormalities may be represented as a spectrum (Table 14.5). Similar, but less pronounced changes occur in postmenopausal women with chronic liver disease. Three-quarters of men with cirrhosis have oligospermia and this is associated with normal plasma concentrations of FSH. The absence of a compensatory rise in FSH has been attributed to a hypothalamic defect that is characteristic of patients with cirrhosis. Absence of testicular atrophy, a normal LH or a normal LH or FSH response to clomifene or gonadotrophin releasing hormone predict recovery of sexual function in alcoholic men who abstain from further drinking.

Hypogonadism is a prominent feature in men with hereditary haemochromatosis. Unlike the situation with other types of cirrhosis, such as that due to alcohol, impotence may occur very early in the disease (and even before the cirrhosis develops). It is probably due to a combination of hypogonadotrophic hypogonadism, concurrent hepatic cirrhosis and diabetic autonomic neuropathy. The pituitary dysfunction, which is probably the most important factor, is caused by selective iron deposition in the gonadotrophin-secreting cells. Plasma concentrations of LH (and prolactin) are subnormal in the majority of men with clinical hypogonadism associated with liver disease.

Changes in liver function during pregnancy

Abnormalities (occasionally marked) in the LFTs may occur during normal pregnancy. These have been attributed, not unreasonably, to changes in the hormonal environment and there is good experimental evidence that oestrogens are involved. Thus, similar abnormalities are also seen in women taking the contraceptive pill or other oestrogenic preparations.

The most pronounced change is a tendency towards cholestasis in the last trimester – intrahepatic cholestasis of pregnancy. The plasma activities of γ GT and ALP rise in late pregnancy but the latter is attributable mainly to placental ALP. In a small percentage of women, however, this tendency is exaggerated and symptoms of jaundice and/or pruritus develop. Although different names are given depending on the

predominant symptom, they probably form part of a spectrum. Signs and symptoms disappear within a few days, or even hours, of delivery but recur with varying degrees of severity in subsequent pregnancies or on exposure to oestrogen containing preparations. Female relatives of patients with intrahepatic cholestasis of pregnancy are at an increased risk of developing the same syndrome. Plasma total bile acid concentration is typically increased 5–10-fold and this may sometimes be the only abnormality.

Severe liver disease leading to acute liver failure is a rare complication of pregnancy. Although these conditions probably represent a clinicopathological continuum, the clinical features and timing during pregnancy, results of laboratory tests and maternal and fetal outcomes allow differentiation between acute fatty liver of pregnancy, the 'haemolysis, elevated liver enzymes and low platelets' (HELLP) syndrome and liver impairment in eclampsia and hyperemesis gravidarum. Liver rupture is a rare complication of fatty liver and the HELLP syndrome, with high mortality. With the widespread introduction of testing for viral hepatitis in pregnancy and biochemical screening, it is now clear that abnormal LFTs can occur in up to 3% of pregnancies. Details of appropriate investigations and management of liver disease in pregnancy are available in specialized texts.

Not surprisingly, the excretory defect in patients with the Dubin–Johnson syndrome (Chapter 13) is worsened by oestrogens, and such patients often first present with, or suffer from, increasing jaundice during pregnancy or on first exposure to the contraceptive pill. Pregnancy and the contraceptive pill are also associated with the development of benign liver tumours and increased blood coagulability leading to hepatic vein thrombosis, but both of these complications are very rare.

Changes in hormonal profiles during or after pregnancy can also precipitate, or exacerbate pre-existing AIH in women. Occasionally, this rare condition presents for the first time during the second or third trimesters of pregnancy or in the immediate postpartum period. More frequently, patients with pre-existing AIH, who have been in treatment-induced remission, experience flares of their disease during or after pregnancy. These flares are unpredictable and AIH patients who become pregnant, therefore, require more frequent monitoring of their serum aminotransferase activities during pregnancy and for up to six months after delivery, so that their immunosuppressive therapy can be increased (or reintroduced) at an early stage to avoid the development of subacute liver failure, which carries a high risk for both mother and fetus. The mechanisms underlying these effects of hormonal changes

on the activity of this disease are not well understood, but it is known that hormones have profound effects on the immune system. Thus, high oestrogen and prolactin concentrations promote mainly a cell-mediated response, while low oestrogen and high progesterone concentrations favour an antibody-mediated pathway.

Glucose intolerance

Most patients with hepatic cirrhosis exhibit intolerance to glucose. This is probably due to insulin resistance at the level of peripheral glucose uptake and glycogen synthesis by muscle. The fasting plasma concentration of insulin is increased because of a combination of decreased hepatic extraction, enhanced secretion and portosystemic shunting.

Hypoglycaemia, presumably because of the vast functional hepatic reserve, is rare in liver disease, being virtually confined to patients with acute liver failure, galactosaemia, fructosaemia or, even more rarely, primary hepatic malignancy. Hypoglycaemia can occur in acute liver failure and can be severe and persistent; for this reason such patients should be routinely maintained on a glucose infusion, particularly if they are being transferred between hospitals.

In patients with type 2 diabetes mellitus and liver disease who require treatment with oral hypoglycaemics, biguanides are best avoided because of the danger of lactic acidosis. Most sulfonylureas are metabolized in the liver, and those with a long plasma half-life should be avoided in patients with liver disease because of the risk of hypoglycaemia owing to accumulation of the drug.

DRUGS AND THE LIVER

Biochemical tests play a crucial role in the recognition and monitoring of adverse effects of therapeutic drugs. An International Consensus Meeting has clarified the terminology best used for reporting and recognizing such reactions, and for inferring causality by the temporal relationship between drug administration and putative hepatic dysfunction. It should be emphasized here that this discussion is not about patients with liver disease, but patients with other conditions who develop some form of liver dysfunction during or after treatment. Table 14.6 lists the definitions and gives some classic examples. In the absence of histological evidence, the term 'drug-induced liver injury' is preferred to 'hepatitis' or

TABLE 14.6 Agreed criteria for classification of drug-induced liver damage

	Laboratory test	ALT/ALP ratio (R)	Examples	
			Acute	Chronic
Hepatocellular	ALT >2×ULN	R >2	Paracetamol, isoniazid	Methyldopa, dantrolene
Cholestatic	ALP >2×ULN	R <2	Anabolic steroids	Phenothiazines
Mixed	ALT and ALP >2×ULN	2 < R < 5	Chlorpromazine	Ajmaline

ULN, upper limit of normal (reference) range.

'cirrhosis'. Almost all the pathological conditions described in this chapter can be caused by drugs, chemicals or toxins. The reader is referred to standard texts for a detailed list of possible associations between liver disease and xenobiotics.

NEOPLASTIC DISEASE OF THE LIVER AND BILIARY TRACT

As with those in other organs, liver tumours are classified as benign or malignant and the latter are subclassified as primary (relatively rare) or secondary (relatively common). Benign liver tumours are of little clinical significance other than as a very occasional association with the use of oral contraceptive preparations and their increasing incidental detection during ultrasound or computerized tomography (CT) imaging. The extent to which incidentally detected space-occupying lesions of the liver should be further investigated is difficult and controversial: most will be benign and of no clinical significance. They should be investigated, however, because of the devastating consequences of missing the occasional malignant tumour that, when detected early, may be amenable to surgical resection. In general, LFTs are normal in subjects with benign lesions, so there is little doubt that further investigation should be undertaken if LFTs are abnormal in any respect.

In the West, most malignant disease of the liver involves metastatic deposits from primary tumours of the gut, pancreas, lung or breast. Abnormal LFTs do not provide any specific information about the primary site of a tumour metastasizing to the liver. On the other hand, metastases will seldom be of sufficient size to be detectable by imaging in individuals with normal LFTs. Measurement of carcinoembryonic antigen (CEA) is of little diagnostic use, but serial measurement after resection of a primary colorectal tumour may provide early evidence of disease recurrence, particularly in the liver.

Hepatocellular carcinoma and α -fetoprotein

Primary liver cancer is relatively uncommon in the West but has been increasing due, in part, to the increased prevalence of chronic hepatitis C. However, worldwide, hepatocellular carcinoma (HCC) is one of the commonest and most rapidly progressive malignant neoplasms. In more than 75% of patients, it arises as a complication of hepatic cirrhosis. Hepatocellular carcinoma is one of the few tumours for which a serological marker, α -fetoprotein (AFP), is clinically useful and widely available, although the sensitivity of AFP is only of the order of 80% as not all tumours secrete this protein. The reference range for AFP is usually quoted as $<10\mu\text{g/L}$, while values of up to $10^7\mu\text{g/L}$ are seen in patients with HCC. Values $>500\mu\text{g/L}$, in an appropriate clinical setting, are virtually diagnostic of HCC, but there is a 'grey area' between 10 and $500\mu\text{g/L}$, where similar values can be seen in patients

with uncomplicated chronic liver disease associated with hepatic regeneration.

Unfortunately, many patients with small and potentially resectable tumours fall into this 'grey area' and, consequently, several attempts to improve the specificity of the test have been made. A steady increase in AFP concentration within the 'grey area' is particularly suggestive of malignant change. Also, there is good evidence that a high percentage of AFP originating from malignant cells is hyperfucosylated, and a routine test for this fraction may become available in the future. Apart from its use in diagnosis of HCC, AFP concentration is widely used to monitor response to therapy, since the concentration falls and rises in relation to the tumour mass.

A rare type of HCC, the fibrolamellar variant, is characterized by occurrence in adolescence, absence of AFP and a rather better prognosis than other types of HCC. The plasma unsaturated B_{12} binding protein (UBBP) and neurotensin concentrations are usually grossly elevated. The diagnosis is, however, made on histological grounds, although measurement of UBBP and neurotensin can be used to monitor treatment.

PARENTERAL NUTRITION

Abnormalities of LFTs are frequently noted in patients receiving parenteral nutrition (PN) and usually arise during or after the second week of therapy. An increase in ALP and aminotransferases to more than twice the upper limit of the reference range occurs in up to 50% of patients, but few actually develop clinical jaundice. The underlying histological lesion is usually steatosis, and there seems little doubt that over-provision of calories and infusate imbalance (particularly when glucose alone rather than glucose and fat is used as the energy source) are involved. Much anxiety has been expressed about the hepatotoxicity of fat emulsions, but this only occurs if very high concentrations are used.

The possibility that changes in LFTs might be attributable to a specific complication of the condition for which the PN is being used must be borne in mind. An additional complication of PN is the development of both gallstone disease and acalculous cholecystitis. In neonates, particularly when premature, the development of PN-related liver disease may limit its use. In such children, the hepatic lesion ranges from mild cholestasis associated with steatosis to severe cholestasis, cirrhosis and liver failure.

BACTERIAL INFECTIONS

Extrahepatic bacteraemia is frequently associated with abnormal LFTs and, particularly in the young, jaundice. The picture is usually one of cholestasis, although in certain specific instances, for example in the toxic shock syndrome (due to *Staphylococcus aureus* infection), the picture may be more hepatic.

Spontaneous bacterial peritonitis may occur in subjects with ascites and liver disease, in the absence of any

TABLE 14.7 Relation of clinical signs and ascitic fluid content in patients with ascites

	Clinical signs of peritonitis	Ascitic fluid analysis		
		PMN count	Gram stain	Culture
Normal/sterile	None	<250	–	Negative
'Bacterascites'	None	<250	±	Positive
Culture-negative neutrocytic or 'probable' peritonitis	Yes	>500		Negative
'Silent' peritonitis'	Minimal ^a	>500	±	Positive
Spontaneous bacterial peritonitis	Yes	>250	±	Positive

^aSome patients have fever or abdominal pain but without signs of peritonism. PMN, polymorphonuclear leukocytes. (Rolando N, Wyke R J. *Gut* 1991: S25–8).

apparent intra-abdominal source. The condition may be clinically silent, the only manifestations being onset of encephalopathy, deterioration in renal function and worsening LFTs. Rapid detection is essential and diagnosis is based on measurement of the polymorphonuclear leukocyte count in ascites and Gram stain of a centrifuged pellet (Table 14.7). The bacterial culture of ascites should be performed by directly injecting freshly drawn ascites into blood culture bottles at the bedside, because transporting ascites to the microbiology laboratory results in a significant reduction in the success of identifying the infecting organism. The responsible organism is usually *Escherichia coli* or other gut-related organisms and, less frequently, pneumococci.

INHERITED METABOLIC DISORDERS INVOLVING THE LIVER

Iron overload and hereditary haemochromatosis

Iron is a toxic metal, and whole body iron balance is tightly regulated at the level of the duodenal mucosa by controlling the amount of iron absorbed from the diet. Iron excretion is limited, and in the absence of a regulated pathway of iron loss, positive iron balance leads to increased tissue concentrations. The term haemochromatosis refers to the group of genetic disorders in which, as a result of excessive absorption of dietary iron and long-term positive iron balance, iron deposition causes tissue damage, particularly to the liver, pancreas, heart, anterior pituitary and joints. Haemosiderosis implies iron overload without tissue damage, often an early stage of iron accumulation, while secondary haemochromatosis occurs in conditions requiring multiple blood transfusions and in some other haematological disorders.

Although previously considered a single gene disorder, it is now known that haemochromatosis can be caused by mutations in several genes that appear to have different functions in iron metabolism. The most common form of

haemochromatosis, found almost exclusively in people of northern European descent, is caused by homozygosity for a low penetrant mutation, C282Y, in the hereditary haemochromatosis gene, *HFE*. This condition affects mainly men and is characterized by the insidious accumulation of iron, with the onset of symptoms and signs of iron overload delayed until the fourth or fifth decades of life. Heterozygotes for the C282Y mutation do not develop iron overload, although minor abnormalities in plasma iron and ferritin concentrations occur in approximately 15% of such individuals. Adult onset haemochromatosis is very rarely associated with mutations in *TfR2*, the gene encoding transferrin receptor 2, a cell surface glycoprotein involved in iron transport and uptake by cells including hepatocytes.

A contrasting group of disorders, in which the iron loading process is very rapid, presenting by the second or third decades and affecting males and females equally, is known as juvenile haemochromatosis. Although liver disease is invariably present, usually as cirrhosis, the clinical presentation in this form of genetic iron overload is with cardiac and endocrine failure. Juvenile haemochromatosis is caused by mutations in *HJV* and *HAMP*, genes that encode, respectively, hemojuvelin and hepcidin. The iron loading process resulting from the digenic inheritance of mutations in *HFE* and *TfR2*, genes normally associated with adult onset haemochromatosis, can be so rapid as to lead to the phenotype of juvenile haemochromatosis. The recent identification of the iron regulatory hormone hepcidin provides a link between *HFE*, *TfR2* and *HJV*, since mutations in these genes, and in *HAMP* itself, cause loss of hepcidin production by the liver. This peptide acts by inhibiting dietary iron absorption and iron release from recycling and storage sites. All these forms of iron overload are characterized by hepcidin deficiency and it appears that the proteins encoded by *HFE*, *TfR2* and *HJV* function as sensors of iron status on the hepatocyte membrane acting upstream of *HAMP*. Genotyping for mutations in *HFE* is now an essential part of diagnosis and screening for haemochromatosis. Mutations in the other genes are so rare that routine genetic analysis for these is not undertaken. A further variant of haemochromatosis is associated with mutations in *ferroportin*, a gene that encodes the hepcidin receptor.

The penetrance of the homozygous C282Y genotype in *HFE*-related haemochromatosis, defined in terms of severe iron overload with tissue damage manifesting as cirrhosis and type 1 diabetes, is low, probably in the order of ~1–2%. Biochemical penetrance, defined as an increase in transferrin saturation of >60% and a minimally raised plasma ferritin concentration, probably occurs in 20–50% of homozygous individuals. Other undetermined genetic loci and possibly environmental factors are likely to determine penetrance, but these are currently uncharacterized.

A clinical diagnosis of haemochromatosis in a severely iron loaded subject may be made on the basis of signs of liver disease, glycosuria and a slate-grey appearance of the skin, but confirmation of the diagnosis is entirely dependent on genetic and laboratory tests. The crucial investigations are plasma iron and ferritin concentration and transferrin saturation. In a patient

with full gene penetrance, the fasting plasma iron concentration is usually $>40\ \mu\text{mol/L}$ (reference range $10\text{--}30\ \mu\text{mol/L}$) and transferrin is usually $>60\%$ saturated. Plasma ferritin concentration, which is approximately proportional to the iron excess, is usually $>1000\ \mu\text{g/L}$ (upper limit of reference range $200\text{--}300\ \mu\text{g/L}$) and starts to rise when hepatic iron stores exceed twice the reference range.

If these tests suggest iron overload and liver function tests are abnormal, a liver biopsy should be undertaken to assess the degree of liver damage. Histological staining for iron (using Perls stain) gives a semiquantitative assessment of the degree of iron overload, but the amount of iron in a biopsy can also be quantitated directly by inductively coupled mass spectrophotometry (ICP-MS). The reference range is up to about $20\ \mu\text{mol/g}$ dry liver weight and $>40\ \mu\text{mol/g}$ dry liver weight is found in patients with haemochromatosis. In less severe disease, when the patient is <40 years of age, plasma ferritin is $<1000\ \mu\text{g/L}$, LFTs are normal and genotyping confirms the homozygous C282Y mutation, then liver biopsy is not indicated as cirrhosis is unlikely to have developed.

Treatment involves weekly venesection until the patient starts to develop iron deficiency anaemia and the plasma ferritin concentration falls to $\sim 50\ \mu\text{g/L}$. Since $500\ \text{mL}$ of blood contains about $250\ \text{mg}$ of iron, the total body iron at the time of diagnosis can be calculated from the total volume removed. In an iron loaded individual, it is usually in the order of $10\text{--}25\ \text{g}$ compared with less than $4\ \text{g}$ in the normal individual. The same figures can be obtained from knowledge of the plasma ferritin concentration, which can also be used to monitor the progress of treatment ($1\ \mu\text{g/L}$ of ferritin is equivalent to approximately $8\ \text{mg}$ of iron). Unless cirrhosis has already occurred, phlebotomy is successful in preventing progressive liver disease. The high risk of development of hepatocellular carcinoma remains in treated patients with cirrhosis. It is also important to screen the patient's relatives for haemochromatosis by *HFE* genotyping so that treatment can be instituted if appropriate before liver damage occurs. Plasma ferritin and transferrin saturation are the most sensitive biochemical tests for presymptomatic detection in population studies.

Wilson disease

Wilson disease is an autosomal recessively inherited disease of copper metabolism with a worldwide prevalence of about 1:30 000 and a carrier rate of 1:90. It leads to excessive deposition of copper in the liver, kidneys, brain, eyes and other tissues. Like haemochromatosis, it is treatable and thus has an importance in liver disease out of proportion to its rarity. Liver disease is the most prominent aspect in about 40% of patients, the remainder exhibiting neurological, psychiatric or haematological complications. The hepatic manifestations vary widely from asymptomatic development of cirrhosis to an acute onset, often accompanied by haemolysis and evidence of renal tubular dysfunction, which may progress to fulminant hepatic failure. In such patients, the finding of a very low serum alkaline phosphatase activity may be

a useful pointer to Wilson disease: the transient marked hypercupraemia is thought to lead to incorporation of copper instead of zinc into the active sites of ALP, reducing its activity. Accumulation of copper in the eyes is associated with the appearance of Kayser–Fleischer rings in the cornea (detectable by slit-lamp examination), although these may not always be present. In the brain, the copper deposition leads to neurological symptoms (see Chapter 36).

The defect in Wilson disease is attributed to mutations in the *ATP7B* gene on chromosome 13 (at the 13q14.3 locus), which cause failure of ATPase-dependent copper excretion. To date, more than 200 mutations have been identified, and most patients are compound heterozygotes. However, in some patients with classic Wilson disease, no mutations in the *ATP7B* gene have been found. Nonetheless, progress of the disease can be effectively arrested by treatment (see below) and, therefore, all first-degree relatives of patients with Wilson disease should be screened, initially by biochemical testing, to identify asymptomatic siblings who might benefit from early therapeutic intervention.

Among those presenting with chronic liver disease, the elevation of aminotransferases is said to be inappropriately low for the degree of hepatic inflammation and ALP is often within the reference range, or low in patients where zinc depletion has occurred. There may be a low-grade haemolytic anaemia leading to mild hyperbilirubinaemia, with an increased risk of pigment gallstones. Copper deposition may lead to renal tubular dysfunction with loss of potassium, glucose, amino acids and phosphate.

Diagnosis

The diagnosis, or exclusion, of Wilson disease is of crucial importance to patients and their relatives. It is seldom as straightforward as it may appear in standard texts. If there is any doubt about the diagnosis, advice should be sought from a specialist centre. The following are characteristic laboratory features (see Table 14.8).

1. **A low plasma caeruloplasmin concentration.** More than 90% of patients have values below the reference range of $200\text{--}400\ \text{mg/L}$. The only exception is when there is very 'active' liver disease, when values may be within the lower part of the reference range. About 10% of asymptomatic carriers will also exhibit a low plasma caeruloplasmin concentration. Caeruloplasmin synthesis is induced by steroids and plasma concentrations are higher in women using the oral contraceptive pill or hormone replacement therapy.
2. **A low plasma copper concentration.** This is usually below the lower limit of the reference range, but similar values can be found in association with other disorders and it therefore lacks diagnostic specificity. Approximately 90% ($10\text{--}25\ \mu\text{mol/L}$) of plasma copper is irreversibly bound to caeruloplasmin. The remainder ($<2\ \mu\text{mol/L}$), termed 'free' copper, is loosely bound to albumin and to amino acids. A 'free' plasma copper concentration $>7\ \mu\text{mol/L}$ is virtually diagnostic of Wilson disease,

TABLE 14.8 Typical laboratory findings in Wilson disease and other conditions in which Wilson disease may enter the differential diagnosis

Test (reference range)	<i>Wilson disease</i>					
	Chronic liver disease	Fulminant hepatic failure	Asymptomatic	Acute hepatic necrosis ^a	Chronic cholestasis ^b	Indian childhood cirrhosis
Liver function tests	Mildly abnormal	Grossly deranged	Normal	Grossly deranged	Cholestatic	Mildly abnormal
Plasma copper						
total (10–30 µmol/L)	Variable	>30	Variable	>15	>25	>25
free (<2 µmol/L)	>10	>10	>2	<2	<2	<2
Plasma caeruloplasmin (200–400 mg/L)	<200	<200 but may be normal	<270	100–270	200–400	200–400
Urinary copper						
basal (0.25–0.75 µmol/24 h)	>2	>15	>1	<15	>1	>0.75
penicillamine stimulated	>20-fold increase	>10-fold increase				
Hepatic copper (1 µmol/g dry weight)	4–20	20–40	1–4	<1	2–15	1–4

^aAny severe hepatitic illness.

^bConditions such as primary biliary cirrhosis and sclerosing cholangitis.

but such values are found in only ~50% of patients. Normal total plasma copper concentrations can often be seen in patients with acute hepatocyte damage owing to release of copper trapped in the liver.

3. **A high tissue copper concentration.** It is clearly important that the diagnosis of Wilson disease is entertained before liver biopsy is undertaken so that an adequate sample may be taken for estimation of liver copper concentration. It is worth taking two cores to ensure that enough material is available, one for routine histology and one for tissue copper estimation, to reduce the risk of a falsely low value due to sampling error. Liver copper concentrations can, however, be raised in cholestatic liver diseases other than Wilson disease, reducing the specificity of the test.

The combination of a very low plasma copper concentration with a very high tissue copper concentration is diagnostic of Wilson disease, especially when Kayser–Fleischer rings are detected. However, not infrequently, either a liver biopsy is contraindicated because of severe coagulopathy or tissue concentrations are not high enough to be diagnostic, so other confirmatory tests may be required. The most reliable of these (and the simplest to perform) is the penicillamine challenge test. This involves collection of two 24 h urine samples for copper estimation before and after administration of D-penicillamine. This is given as two oral 500 mg doses, one at the start and one in the middle of the second 24 h period. A 10–20-fold increase in the 24 h urinary copper excretion after penicillamine administration is virtually diagnostic of Wilson disease. Advances in DNA sequencing techniques have meant that molecular diagnosis is becoming more readily available and is particularly helpful in diagnosing asymptomatic siblings once a family mutation in the *ATP7B* gene has been identified.

Long-term management of hepatic Wilson disease

Wilson disease is treated with penicillamine, which leads to a gradual improvement in standard LFTs. The rationale for the use of penicillamine was to increase urinary copper excretion and thereby reduce body copper. Despite the undoubted efficacy of the drug in Wilson disease, it now appears that liver copper concentrations do not change significantly and that the mechanism of action of penicillamine is considerably more complex than originally thought.

Penicillamine therapy is associated with a number of adverse drug reactions, including skin rashes, proteinuria, leukopenia and thrombocytopenia. It is particularly important to test the urine for protein during penicillamine treatment, because the drug can precipitate a number of immunologically mediated conditions including a form of nephrotic syndrome and Goodpasture syndrome. Treatment must be life-long, and even short interruptions may precipitate fatal hepatic failure. Trientine hydrochloride is an effective second-line treatment in patients who are intolerant of penicillamine. Since some may find penicillamine tablets unpleasant or inconvenient to take, and some children will never have had, or will not be able to remember having had, any symptoms of the disease, compliance may be a problem. Zinc has been used as monotherapy in children with Wilson disease; it acts as an inducer of gut mucosal metallothionein, which binds copper and leads to excretion in the stools. The best indication that an effective dose is being taken is said to be that the plasma copper concentration remains low. The use of zinc monotherapy in adults is controversial and in any event should only be introduced following treatment with either penicillamine or trientine for some years. Life-long monitoring with liver function tests and urinary copper measurements is mandatory in Wilson disease.

Indian childhood cirrhosis

This condition, which affects young children, was thought to be confined to the Indian subcontinent. The liver is grossly fibrotic and concentrations of tissue copper are in the same range as those seen in Wilson disease (see Table 14.8). Variants are now seen in other geographical areas and it appears that different mechanisms can lead to the same end-stage chronic liver disease. The Indian subcontinent variety is probably a childhood cirrhosis that requires a high environmental exposure to copper for full expression, as supplied by excess dietary copper (from milk stored in copper vessels). In the USA, a cholestatic process seems to underlie a similar form of cirrhosis with high concentrations of tissue copper. A genetically deleterious excretory defect is likely to be present. Arsenic ingestion has also been implicated. Penicillamine may be effective if given before decompensation develops.

α_1 -Antitrypsin deficiency

Deficiency of this glycoprotein, which is an enzyme that acts as an anti-proteinase predominantly inhibiting elastase, is associated with both pulmonary and liver disease. The condition is inherited co-dominantly and, in the classic form of the disease, histological examination of liver biopsy specimens reveals characteristic PAS-positive, diastase-resistant globules within the portal tracts. These globules represent accumulation of the defective protein within the endoplasmic reticulum. The mechanism of liver cell damage, which often progresses to cirrhosis, is ill understood.

The gene encoding α_1 -antitrypsin (α_1 -AT) is located on chromosome 14 (at 14q32.1). More than 90 genetic variants have been described, most of which are exceedingly rare and are associated with production of normal amounts of functional enzyme. The different variants (phenotypes) can be identified by their mobility on isoelectric focusing. The conventional nomenclature describes the various phenotypes by the letters 'Pi' (protease inhibitor) followed by an alphabetical code relating to their mobilities. The commonest (normal) phenotype is designated 'M' (medium) and is further subdivided into at least eight subtypes (M1, M2, M3 etc.). Of the mutations that result in α_1 AT deficiencies, the two commonest are designated 'S' (slow) and 'Z' (ultraslow), with homozygotes being described as PiSS or PiZZ, compound heterozygotes with defective phenotypes as PiSZ, and heterozygotes as, for example, PiMS or PiMZ. The S phenotype is mainly associated with lung disease, while the Z phenotype is strongly associated with both lung and liver disease. Other rare phenotypes are also associated with liver disease (S_{Iiyama}, M_{Malton}, Pi_{King}) (see also Chapter 43).

Individuals with defective phenotypes may present (or be identified incidentally) at any time from infancy to old age. The most important period is during the first few weeks of life, with about 25% of cases of the neonatal hepatitis syndrome being due to this condition. The association with liver disease is very variable. Some children, particularly those who do not develop symptoms

in infancy, may never develop liver disease, while others may progress rapidly to cirrhosis and death from liver failure. The Z allele causes a Glu342Lys substitution in the protein that results in the formation of polymers that are retained in the endoplasmic reticulum of hepatocytes. These polymers are degraded via several pathways, including the ubiquitin–proteasome and the autophagy–lysosome systems. There is significant interindividual variation in the capacity of these systems, which might account for the variation in the extent of liver disease in affected individuals. The accumulation of polymers of α_1 -AT in the liver means that only 10–15% of the α_1 -AT molecules produced reach the circulation. This forms the basis for the initial screening test as plasma α_1 -AT concentrations are low (typically 0.1–0.2 g/L, normal 1.0–2.0 g/L). α_1 -Antitrypsin is a positive acute phase reactant and upregulation of α_1 -AT synthesis during an inflammatory response or acute hepatocellular damage can cause release of the accumulated polymers, increasing plasma concentrations to within the reference range. Strategies for investigation of possible α_1 -AT deficiency in these situations should include phenotyping and/or genotyping.

The hepatic porphyrias

These are so called because the major organ in which excessive production of the porphyrins or their precursors takes place is the liver (see Chapter 28). However, the four types (acute intermittent porphyria, hereditary coproporphyrin, variegate porphyria and 5-aminolaevulinic acid dehydratase deficiency porphyria (Doss) porphyria) do not have any hepatic manifestations other than the occasional mild rise in plasma aminotransferase activity and an increased incidence of hepatocellular carcinoma in patients with acute intermittent porphyria. Porphyria cutanea tarda has already been described (p. 258). Asymptomatic porphyrinuria (mainly due to coproporphyrin) is common in many acute liver diseases and hepatic tumours. In erythropoietic porphyria, the liver plays a minimal role in protoporphyrin overproduction, but hepatic complications may occasionally arise: these comprise gallstones and cirrhosis associated with massive protoporphyrin deposition.

Cystic fibrosis

A wide spectrum of hepatobiliary complications of cystic fibrosis is becoming increasingly recognized as larger numbers of children survive to adulthood as a result of improvements in the treatment of the pulmonary complications of the disorder. In infancy, cystic fibrosis may present as neonatal cholestasis, often in association with meconium ileus. Older children may develop a fatty liver syndrome, but the major problem is a focal biliary cirrhosis that progresses to a generalized cirrhosis with all the complications as described previously. Extrahepatic biliary disease, with stricture of the distal common bile duct, is also frequent, although this, of itself, is probably not directly involved in the pathogenesis of the liver disease. Laboratory investigation is important in the diagnosis of

this condition (see Chapter 25), but the LFTs are entirely non-specific and serve no more than to indicate that liver abnormalities are present. Liver function tests have been used to monitor response to therapy with ursodeoxycholic acid.

Other inherited metabolic diseases

Diagnosis, management and genetic counselling in children with inborn errors of metabolism form a large part of paediatric hepatology. The disorders also illustrate many aspects of applied intermediary metabolism. Most are rare and few laboratories will develop wide experience. Here, the more common types are described briefly, with more detailed accounts in the Further reading list at the end of the chapter.

Tyrosinaemia

This is an autosomal recessive condition, of which there are two types. Type I is due to deficiency of fumarylacetoacetate hydrolase (Fig. 14.3) and is the only genetically mediated disorder of amino acid metabolism that has so far been identified as a significant cause of liver disease. Type II is due to deficiency of tyrosine aminotransferase and is not associated with liver damage.

Type I tyrosinaemia usually presents at the age of 2–6 months with vomiting, diarrhoea, oedema, ascites, hepatosplenomegaly, hypoglycaemia and failure to

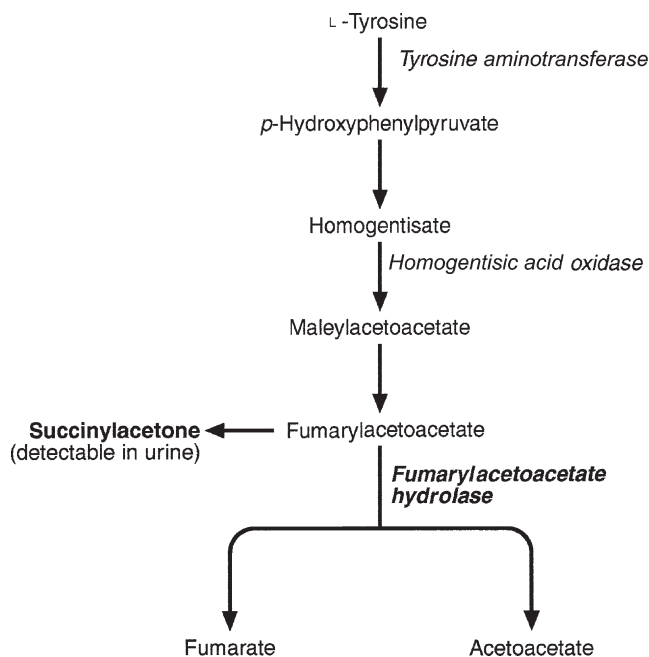


FIGURE 14.3 ■ Intermediary metabolism of tyrosine. Classic tyrosinaemia (type I or tyrosinosis) is caused by a defect in fumarylacetoacetate hydrolase activity, but defects in other parts of the pathway are also recognized including alkaptonuria, in which there is deficiency of homogentisic acid oxidase. There are increases of all precursors (including tyrosine) in the urine, but the diagnostic test is for succinylacetone which, by reacting with sulphhydryl groups, may in part be responsible for the tissue damage.

thrive. There is a marked bleeding diathesis owing to prothrombin deficiency. Other characteristic features include systemic acidosis and renal tubular dysfunction leading to hypophosphataemic rickets. Rapid and accurate diagnosis is essential since dietary restriction of phenylalanine, tyrosine and methionine may improve the prognosis in this otherwise fatal disease. A more chronic form of the disease may present later in childhood with cirrhosis and a high incidence of hepatocellular carcinoma, although AFP concentrations may be grossly elevated even in the absence of a tumour.

The characteristic laboratory findings are a marked amino acidemia, with tyrosine, phenylalanine and methionine concentrations being particularly elevated. There is also moderate derangement of LFTs, a raised INR and a markedly raised AFP. However, the diagnostic finding is the presence of succinylacetone in urine (or amniotic fluid, for prenatal diagnosis), usually with elevated urinary 5-aminolaevulinate (owing to the inhibition of porphobilinogen synthetase by the succinylacetone). Liver transplantation is being increasingly used in both forms of the disease. Treatment with nitisone may be beneficial in slowing progression of the condition (see Chapter 24).

Galactosaemia

Galactose is generated by intestinal hydrolysis of lactose and is converted in the liver to glucose. Classic galactosaemia (see Chapter 24) is an autosomal recessive condition that is due to deficiency of the enzyme galactose 1-phosphate uridylyltransferase (GALT), resulting in accumulation of galactose, galactitol (via an aldoreductase) and galactose 1-phosphate in all body tissues (Fig. 14.4). The disease usually becomes evident soon after birth when milk is introduced into the diet. It varies widely in severity but, in the majority, the LFTs are grossly deranged, with elevated plasma aminotransferases and a raised INR, indicating hepatocellular necrosis. There is a severe unconjugated jaundice due to haemolytic anaemia, accompanied by amino aciduria and proteinuria, indicating renal damage. Often, the condition rapidly progresses to liver failure, with bleeding and ascites. The mechanisms underlying the liver damage are poorly understood, but may be related to cellular ATP depletion owing to accumulation of galactose 1-phosphate in cells, which may restrict the availability of phosphate for formation of high-energy bonds. A less acute presentation may be with a chronic liver disease, cataracts and learning difficulties. Children with this condition have an increased risk of septicaemia.

Galactosaemia should be suspected if sugars other than glucose are detected in urine. In this context, it should be noted that conventional urine stick tests detect only glucose and not other reducing sugars. The diagnosis requires the demonstration of low or absent GALT enzyme activity in erythrocytes and confirmation by genetic testing (see Chapter 24). Treatment is by withdrawal of lactose and galactose from the diet, which can be monitored by measurement of galactose-1-phosphate in blood samples.

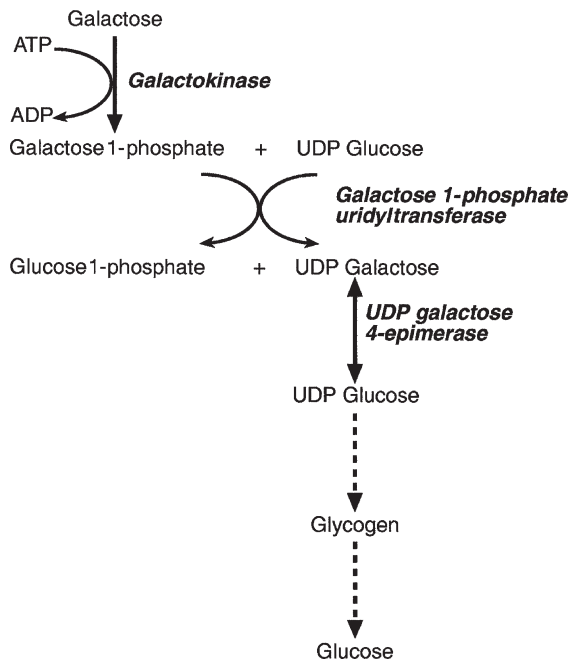


FIGURE 14.4 ■ Metabolism of galactose. The usual lesion in classic galactosaemia is a decreased activity of galactose 1-phosphate uridylyltransferase, but the other enzymes listed may also be involved in some cases. Note that the epimerase activity is normal in most affected individuals so that even on a galactose-free diet, galactosaemic individuals generate enough galactose for normal development.

Fructose intolerance

Three genetic defects in fructose intolerance have been identified: essential fructosaemia (hepatic fructokinase deficiency), hereditary fructose intolerance (fructose 1-phosphatase deficiency), and fructose 1,6-diphosphatase deficiency. All three are relatively benign conditions if fructose intake is avoided or substantially reduced. However, with high fructose intake, both hereditary fructose intolerance and fructose 1,6-diphosphatase deficiency can lead to severe liver and kidney damage. The mechanisms involved are not well understood but, by analogy with galactosaemia, the hepatic injury may be due to intracellular accumulation of phosphorylated sugars reducing availability of phosphate for ATP formation. The liver damage is reflected in grossly deranged LFTs, with markedly elevated aminotransferase activities, a conjugated hyperbilirubinaemia, hypoalbuminaemia and a raised INR, accompanied by hypokalaemia, hypophosphataemia, profound hypoglycaemia and thrombocytopenia. The renal damage leads to aminoaciduria and proteinuria.

Fructose 1,6-diphosphatase deficiency usually has less severe effects on the liver. Plasma aminotransferase activities tend to be mildly or moderately elevated and the bilirubin concentration is only mildly raised. The prominent features are hypoglycaemia, lactic acidemia, aminoacidemia (particularly alanine and glutamine) and a marked ketoacidosis. Diagnoses for all three conditions require demonstration of the deficiency of the particular enzyme in liver biopsy specimens, or genetic testing.

The sphingolipidoses and Niemann–Pick disease type C

The sphingolipidoses are recessively inherited lysosomal disorders in which there is a deficiency of a specific lysosomal hydrolytic enzyme and consequent deposition of complex lipids in various tissues. Type C Niemann–Pick disease is a relatively common inherited cause of liver disease in the UK. In most patients, the disease presents with a fatal neonatal hepatitis, but among those who survive, the liver disease becomes less prominent but death from progressive neurological damage is inevitable.

Glycogen storage diseases

There are 11 distinct types that comprise this group of autosomal recessive diseases, in which enzyme defects result in deficient mobilization of glycogen, deposition of abnormal forms in tissues and very variable prognosis. Types V and VII do not cause liver disease. In the remainder, which can affect the liver, clinical features are usually apparent in childhood and include gross hepatomegaly associated with hypoglycaemic attacks and growth failure. Descriptions of the biochemical lesions, specific clinical features and treatment of all the glycogen storage diseases are beyond the scope of this chapter and the reader is referred to specialist texts. Only type IV is consistently associated with cirrhosis. Liver function tests are abnormal in most types, particularly types III and IV, but jaundice is very rare. Hypoglycaemia is common to all and hyperuricaemia is prominent in types I, III and IX. Among those that affect the liver, the specific diagnoses in types II, III, IV and VI can be established by demonstrating the enzyme deficiencies in leukocytes, but liver biopsies are required for type I. Management revolves around dietary measures to maintain blood glucose concentrations. Liver transplantation is appropriate in some types, particularly types I and III, which can progress to hepatic adenomata and hepatocellular carcinoma.

LIVER TRANSPLANTATION

Liver transplantation is a rapidly expanding part of clinical hepatology. Improvement in survival figures over the last decade has largely been due to better defined indications, improvement in immunosuppressive regimens and organ preservation and improved surgical techniques. The main indication is end-stage chronic liver disease but, as already mentioned, the procedure is being increasingly used for acute liver failure and, in children, for treatment of certain inherited metabolic defects. Although the procedure is now routine in many centres, extensive preoperative assessment is still required and major complications are common. In any event, a liver transplantation programme will impose a heavy load on a clinical pathology laboratory. It is beyond the scope of this chapter to discuss all the problems that may arise, but they are summarized below.

Preoperative assessment

Selection of patients who are likely to benefit from a liver graft requires careful clinical judgement based both on the prognosis and on the quality of life with regard to the primary illness. The criteria are complex and depend on the aetiology of the liver failure and whether it is acute or is a consequence of chronic liver disease. The overall condition of the patient to withstand such a major surgical procedure needs to be carefully assessed. Thus, haemodynamic and respiratory status, patency of major blood vessels and concomitant diseases are essential considerations. The potential effects of the immunosuppressive drug therapy (to prevent graft rejection) on disease recurrence (e.g. in patients transplanted for malignant disorders or viral hepatitis), or on other complications such as pre-existing renal disease also need to be considered. ABO blood group compatibility is required but, in contrast to other solid organ grafting, pre-transplant tissue typing of recipients for human leukocyte antigen (HLA) matching is not essential, except with combined transplants (e.g. liver and kidney), because the liver appears to be an 'immunologically privileged' organ, and donor-recipient incompatibility leading to hyperacute rejection is a rare event (see below).

The immediate postoperative period

The quality of the functioning of the graft immediately after transplantation is a major determinant of the future clinical course and is consequently of great importance. Primary non-function of the graft, that is, a liver that does not appear to work satisfactorily from the outset, is usually due to the poor quality of the donor organ. Hyperacute rejection has been described but is a very rare cause of immediate graft failure. Its differentiation from primary non-function is difficult, but may be suspected if the graft appears to function during the first 6–24 h only to then fail.

The standard LFTs have little value during the immediate postoperative period because the patient's blood will have been extensively diluted by multiple transfusions during the operation. Plasma aminotransferase activities on the second post-operative day are more reliable and are almost always elevated. This represents so-called 'preservation injury' to the donor organ acquired during its removal from the donor, and subsequent preservation and transportation prior to transplantation. Thereafter, a progressive fall in the aminotransferase activities indicates good graft function, while persistently high values suggest that the graft will not function well. Other parameters that are essential for monitoring graft function postoperatively are the blood hydrogen ion concentration (pH), glucose and coagulation factors. Hyperglycaemia is almost always present during the very early postoperative phase but, with normal graft function, the blood glucose usually normalizes over the following 24–48 h. Persistence of the hyperglycaemia, despite increasing doses of insulin, indicates poor graft function.

Acute cellular rejection occurs in 50–70% of patients during the first three weeks after transplantation,

and requires rapid diagnosis and adjustment of immunosuppressive treatment to preserve the graft. It is manifest by an often abrupt increase in plasma aminotransferase activities and bilirubin concentrations. An increasing plasma bilirubin concentration seems to be the more sensitive indicator, but rising aminotransferases appear to be more specific. However, liver histology is required to confirm acute rejection. Some degree of renal impairment is often evident during the first 2–3 days, even if not present before the operation. Serious derangement of the biochemical liver tests, particularly in conjunction with evidence of a coagulopathy, is highly suggestive of interruption to the vascular supply to the graft.

Intermediate follow-up

Frequent monitoring of LFTs is required during the months after a liver transplant to detect signs of late acute or chronic rejection, along with frequent microbiological surveillance for evidence of infections. Late acute rejection (occurring >1 month post-transplantation) is diagnosed as for early acute rejection. Chronic rejection is suspected if the LFTs were previously more or less normal but then begin to become deranged. In this situation, the aminotransferases and INR may be only mildly elevated, but there is a progressive increase in the serum bilirubin and other cholestatic indicators. These changes can also be seen in fungal, bacterial and viral infections, to which the immunosuppressed patient is particularly susceptible, but can usually be differentiated from rejection episodes on clinical and histological criteria together with microbiological investigations.

Patients are generally metabolically unstable for some months after transplantation and require careful adjustment of their immunosuppression. The principal immunosuppressive drugs currently used are ciclosporin, tacrolimus, mycophenolate mofetil and sirolimus, alone or in various combinations, sometimes with glucocorticoids and/or azathioprine. Several of these agents are nephrotoxic and potentially neurotoxic at high doses. Their metabolism is affected by a number of factors, including graft function, other drugs (e.g. antibiotics) that may be required and individual pharmacogenetic variability in their metabolism that can affect the blood concentrations at given doses. Thus, frequent monitoring (initially, two or three times per week) of the concentrations of these drugs in the blood is required during the months following transplantation in order to maintain them within the fairly narrow therapeutic ranges that provide adequate immunosuppression with minimal toxicity.

Long-term monitoring

From the laboratory standpoint, long-term monitoring of liver transplant recipients is relatively straightforward, unless complications arise. Because of the tenuous arterial supply of the common bile duct, biliary tract complications are frequent, but the laboratory features are no different from those seen in patients

with other forms of biliary disease. By one year, most patients are well stabilized and require only three-monthly check-ups with routine LFTs, tests of renal function and haematological investigations. By this time also, they are usually well established on their maintenance immunosuppressive regimens and much less frequent monitoring of their blood drug concentrations is required.

Late graft loss, occurring more than one year after transplantation, is relatively rare and, today, many such patients can be rescued with a second transplant. The commonest causes of late graft failure are uncontrollable chronic rejection and recurrence of the primary disease. Thus, in patients transplanted for chronic viral hepatitis C, there is almost universal reinfection of the graft and consideration needs to be given to pre- and post-transplant antiviral therapy. In the case of patients transplanted for chronic hepatitis B, reinfection can be successfully prevented with regular immunoglobulin administration. The use of newly introduced antiviral agents is also under investigation. In those transplanted for hepatic malignancies, tumour recurrence is not infrequent and, as in other solid organ transplantation, there is an increased risk of non-hepatic malignancies, particularly lymphomas, arising either de novo or through reactivation by immunosuppression of extra-hepatic tumours that have previously been apparently successfully treated. Additionally, as the numbers of long-term survivors rises, recurrence of PBC and AIH (in those transplanted for these conditions) is being seen with increasing frequency.

CONCLUSION

Liver disease is best classified according to its aetiology, qualified by the nature of the resulting pathological change. There are many aetiological factors, including viruses and other infective agents, drugs and toxins, metabolic diseases and autoimmune processes. The range of pathological responses is limited; the most frequent are hepatitis, cirrhosis and cholestasis.

Hepatitis can be acute or chronic: chronic hepatitis can lead to cirrhosis and cholestasis can be a feature of both acute and chronic liver disease. Liver disease frequently has extrahepatic manifestations, reflecting the central role of the liver in metabolism.

Liver transplantation is increasingly being employed in the management of many hitherto fatal liver diseases.

Further reading

- Bircher J, Benhamou JP, McIntyre N et al. editors. Oxford textbook of clinical hepatology. 2nd ed. Oxford: Oxford University Press; 1999.
A comprehensive reference book.
- Criteria for drug-induced liver disorders. Report of an International Consensus Meeting. *J Hepatol* 1990;11:272–6.
Provides a classification of drug-induced hepatic dysfunction and for the inference of causality in this common clinical problem.
- Gleeson D, Heneghan MA. British Society of Gastroenterology guidelines for management of autoimmune hepatitis. *Gut* 2011;60:1611–29.
Recent management guidelines.
- Hannuksela ML, Liisanantti MK, Nissinen AE et al. Biochemical markers of alcoholism. *Clin Chem Lab Med* 2007;45:953–61.
Overview of surrogate markers of alcoholism.
- Joshi D, Heneghan MA, Norris SM et al. Management and outcome of Liver disease in pregnancy. *Lancet* 2010;375:594–605.
Authoritative study of pregnancy in autoimmune hepatitis.
- LaRusso NF, Shneider BL, Black D et al. Primary sclerosing cholangitis: summary of a workshop. *Hepatology* 2006;44:746–64.
Overview of the pathogenesis and management of primary sclerosing cholangitis.
- (Chairs) Ludwig J, McFarlane IG, Rakela J. International Working Party Report: Terminology of chronic hepatitis. *Am J Gastroenterol* 1995;90:181–9.
Detailed recommendations on terminology of chronic hepatitis, which have been adopted and are now used internationally.
- McFarlane IG, Heneghan MA. Autoimmunity and the female liver. *Hepatology* 2004;28:171–6.
Comprehensive review of what is known about hormonal effects on the immune system and how these relate to development of autoimmune liver disease in women.
- McFarlane I, Bomford A, Sherwood R. Liver disease and laboratory medicine. London: ACB Venture Publications; 2000.
A comprehensive account of the use of biochemical, immunological and other laboratory tests in the diagnosis and management of liver disease.
- Merion RM. Current status and future of liver transplantation. *Semin Liver Dis* 2010;30:411–21.
Review of liver transplantation from a medical perspective.
- Mitchell GA, Grompe M, Lambert M et al. Hypertyrosinaemia. In: Scriver CR, Beaudet AL, Sly WS et al. editors. The metabolic and molecular bases of inherited disease. 8th ed. New York: McGraw-Hill; 2001. p. 1777–805.
A detailed account of tyrosinaemia; other chapters in this book cover other metabolic diseases that affect the liver, for example galactosaemia, glycogen storage disease etc.
- Moore KP, Aithal GP. Guidelines on the management of ascites in cirrhosis. *Gut* 2006;55(Suppl. VI):vi1–12.
Review of the pathogenesis and management of a major complication of cirrhosis.
- Mowat AP. Liver disorders in childhood. 3rd ed. London: Butterworth; 1994.
The standard work on paediatric liver disease.
- Pietrangolo A. Molecular insights into the pathogenesis of hereditary haemochromatosis. *Gut* 2006;55:564–8.
Overview of the expanding pool of genes that cause haemochromatosis.
- Tilg H, Hotamiligil GS. Nonalcoholic fatty liver disease: cytokine-adipokine interplay and regulation of insulin resistance. *Gastroenterology* 2006;131:934–45.
Review of the pathogenesis of non-alcoholic fatty liver disease.
- Walsham NE, Sherwood RA. Ethyl glucuronide. *Ann Clin Biochem* 2012;49:110–7.
Review of use of this marker to detect recent heavy alcohol use.

Glucose metabolism and the pathophysiology of diabetes mellitus

David B. Wile • John P.H. Wilding

CHAPTER OUTLINE

PHYSIOLOGY AND PATHOPHYSIOLOGY OF GLUCOSE HOMOEOSTASIS 273

Introduction: the maintenance of normoglycaemia 273

Normal glucose metabolism 274

Insulin 278

Insulin-like growth factors and their receptors 281

CLASSIFICATION AND DIAGNOSIS OF DIABETES MELLITUS 282

Introduction 282

Definitions 282

Type 1 diabetes mellitus 283

Type 2 diabetes mellitus 286

Gestational diabetes mellitus 291

Maturity onset diabetes of the young (MODY) 292

Secondary diabetes 293

ENDOCRINE ASSOCIATIONS WITH DIABETES 295

DIABETES, NUTRITION AND GROWTH 296

MECHANISMS OF DIABETIC TISSUE DAMAGE 296

Introduction 296

Pathogenesis 296

Other aspects of diabetic tissue damage 297

CONDITIONS ASSOCIATED WITH INADEQUATELY CONTROLLED DIABETES MELLITUS 298

BIOCHEMICAL MEASUREMENTS IN DIABETES MELLITUS 299

Glucose measurements 299

Testing for ketones 299

Oral glucose tolerance test 300

Tests of recent glycaemic control 300

Screening for diabetes 301

Tests for insulin resistance 301

Research investigations 303

CONCLUSION 303

PHYSIOLOGY AND PATHOPHYSIOLOGY OF GLUCOSE HOMOEOSTASIS

Introduction: the maintenance of normoglycaemia

Blood glucose concentrations are maintained within very close limits in healthy people. Any given individual has a very strictly maintained postabsorptive (e.g. fasted overnight) blood glucose concentration of 4.5–5.2 mmol/L, with intraindividual coefficients of variation (CV) of only 1–2%. Interindividual CV (assuming similar times since previous meal, levels of activity, composition of previous meal etc.) is less than 5%, so that a fasting glucose of 6.0 mmol/L is usually 4–5 standard deviations away from the mean in most populations. Glucose concentrations in healthy people increase after meals, but typical meals will not raise blood glucose above 8–10 mmol/L and normoglycaemia is usually restored within 2–4 h. Falls in glucose concentration may be produced by severe, sudden unaccustomed exercise or prolonged fasting (or both), by the pathological conditions

discussed in Chapter 17 and by pharmacological means, but are not typically encountered in healthy adults on a daily basis.

The strictness of glucoregulation is remarkable when compared with the relative laxity of regulation of the concentrations of other circulating metabolic fuels such as ketone bodies and non-esterified fatty acids (NEFA, also known as free fatty acids, FFA). The reason for strict avoidance of low blood glucose concentrations is readily apparent in terms of the avoidance of the neurological and other consequences of hypoglycaemia, and it may be no accident that of all the hormones known to influence blood glucose concentration, insulin is the only one able to lower it. The threshold for the onset of detectable neuroglycopenia is of the order of 3.0–3.5 mmol/L, and it is thus appropriate that counter-regulatory mechanisms are set to respond to maintain glycaemia comfortably above this level.

The reason for the strict avoidance of hyperglycaemia is less immediately apparent. Symptoms of hyperglycaemia are florid (in subjects used to relative normoglycaemia) at blood glucose concentrations of >12–13 mmol/L and

may become apparent at concentrations <10 mmol/L. Such concentrations may be seen, for example following a myocardial infarction, so-called 'stress hyperglycaemia' owing to the effect of catecholamine release, and are usually transient. Stress hyperglycaemia also occurs when plasma cytokine concentrations rise, for example in sepsis, with an increase in oxidative metabolism. (The severe metabolic consequences of hyperglycaemia at concentrations usually >20 mmol/L are discussed in the section on diabetic emergencies in Chapter 16.) In contrast, mild hyperglycaemia (glucose 6–9 mmol/L) is usually asymptomatic. The only obvious teleological value of preventing mild hyperglycaemia is the avoidance of the consequences of chronic hyperglycaemia, usually referred to as 'long-term diabetic complications' or 'diabetic tissue damage' (p. 296 and 322), and the increased susceptibility to infection that may occur acutely.

The mechanisms for regulation of normoglycaemia are summarized in **Box 15.1**, which lists the counter-regulatory mechanisms responsible for preventing hypoglycaemia, and **Figure 15.1**, which outlines the main sites of insulin action relevant to prevention of hyperglycaemia.

Normal glucose metabolism

Circulating glucose derives from three main sources:

- the gut, as the result of hydrolysis or hepatic conversion of a variety of ingested carbohydrates
- hepatic and some other glycogen stores (glycogenolysis)
- new synthesis from precursors (gluconeogenesis). Gluconeogenesis takes place in the liver (~75–90%) and kidneys (~10–25%) from glucogenic amino acids, especially alanine, and from glycerol, lactate and pyruvate. Principally it involves the synthesis of a 6-carbon compound from various 3-carbon skeletons. Partial oxidative metabolism in tissues like adipocytes and muscle provides lactate and pyruvate but they can also be donated by red blood cells, in which anaerobic metabolism occurs in the absence of the enzymes of the Krebs (tricarboxylic acid) cycle. Triacylglycerol (triglyceride) released from adipocytes during lipolysis provides glycerol.

Regulation of gluconeogenesis can be via changes in the amount of substrate arriving at the liver or by the

amount that the liver itself extracts, as well as by regulation within the liver. This process is subject to direct hormonal modulation by insulin, glucagon and catecholamines and indirectly by other hormones.

Hepatic glucose output is ~2.0 mg/kg body weight/min in the resting postabsorptive state, or 200–300 g during the average 24 h period (varying with the availability of glucose from food and with the body's requirements during exercise). Plasma glucose concentration is determined by the balance between glucose influx into the circulation (principally from food and hepatic glucose production) and peripheral glucose clearance.

Glucose is stored as glycogen, a 70 kg man typically having a total of 700–1000 g (hydrated) glycogen. Most of this is in the skeletal muscle (400–600 g), and liver (60–125 g), with lesser amounts in other tissues. Glycogen is synthesized from both glucose and the gluconeogenic substrates (see above). Glycogen is a polymeric storage form of glucose. Extension of the polymer by the addition of glucose 6-phosphate subunits is catalysed by glycogen synthase, which is the main regulatory enzyme for glycogen synthesis. This enzyme is itself subject to regulation by a dephosphorylation/phosphorylation cycle controlled by a cyclic AMP dependent protein kinase. Glycogen in skeletal muscle can provide fuel for muscle but does not provide a source of glucose for release into the circulation.

Glucose provides approximately 40–60% (on a typical diet in the developed world) of the total fuel expended by the body during a 24 h period. It provides almost all the energy for the central nervous system (especially in well-nourished subjects, although sustained fasting allows cerebral ketone body utilization). During high-intensity exercise and during the 4–6 h postprandial period, glucose is the predominant fuel for the whole body. Glucose is the most efficient fuel for oxidation in terms of the liberation of energy (112.2 kcal or 6 moles ATP per mole of oxygen consumed). Many tissues can use ketone bodies, fatty acids or glucose for their basal metabolism, switching between these different fuels depending upon their availability in the circulation.

Glucose is fully oxidized to carbon dioxide and water in skeletal muscle, the brain and the liver. The brain accounts for most of the glucose oxidized in the fasting

BOX 15.1 Mechanisms preventing/reversing hypoglycaemia

Adrenergic/sympathetic response

- Promotes glycogenolysis, gluconeogenesis and increased glucose output by the liver; reduces glucose clearance by skeletal muscle and adipose tissue
- Promotes lipolysis to provide alternative fuel source(s)
- Inhibits insulin secretion

'Counter-regulatory hormones'

- Glucagon: promotes glycogenolysis, gluconeogenesis and thus increased glucose output by the liver. May increase hepatic ketone body production

- Cortisol: promotes glycogenolysis, gluconeogenesis and increased glucose output by the liver; reduces glucose clearance by skeletal muscle and adipose tissue
- Growth hormone: promotes hepatic glycogenolysis and increased hepatic glucose output; promotes lipolysis

Other mechanisms

- Insulin secretion inhibited
- Feelings of hunger promote eating
- Hypoglycaemia per se stimulates hepatic glucose output

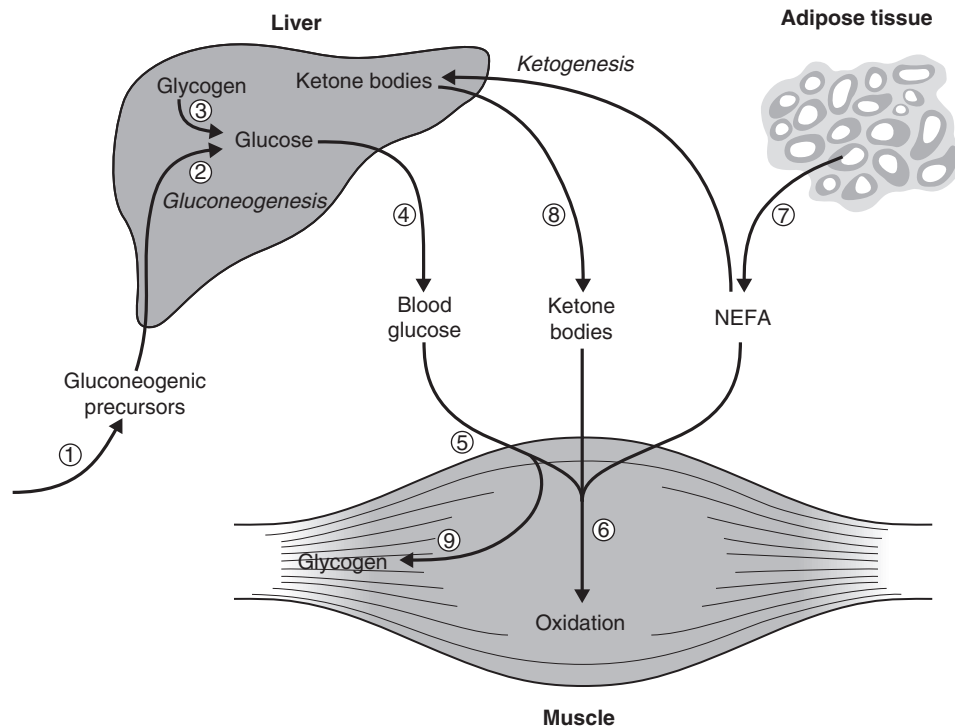


FIGURE 15.1 ■ The principal antihyperglycaemic actions of insulin. Insulin ① reduces the production of gluconeogenic precursors such as glycerol, alanine and lactate, ② reduces activity of hepatic gluconeogenic enzymes and ③ increases hepatic glycogenolysis to glucose. These effects contribute to ④ reduced hepatic glucose output. Insulin ⑤ increases cellular glucose uptake mediated by insulin-sensitive glucose transporters (GLUT4) and ⑥ reduces competition for glucose oxidation by alternative fuels ('Randle effect'). The reduction of competing fuels involves ⑦ inhibition of NEFA release from adipose tissue and ⑧ reduced hepatic ketogenesis. ⑨ Insulin promotes glucose storage as glycogen.

state (100–125 g/24h). In the resting, fasted state, skeletal muscle takes up 10–20% of hepatic glucose output: this is not all oxidized but can be converted to lactate, pyruvate, glycerol or the carbon skeletons of amino acids, much of which subsequently return to the liver and act as precursors for gluconeogenesis. Fatty acids (or their partial oxidation products, ketone bodies) are the prime fuel of resting muscle, heart and liver. Other tissues such as red blood cells, skin, adipose tissue and the renal medulla derive energy from glycolysis to lactate and pyruvate, even in the resting state. Glycolysis to lactate is an anaerobic process to which many cells may resort when faced with hypoxia, for example skeletal muscle during high-intensity exercise. The brain can adapt to the utilization of ketone bodies as a source of fuel during a prolonged fast, but this is far from being an instantaneous process, and the more immediate response depends upon increasing glucose supply as a result of the action of the counter-regulatory hormones glucagon, cortisol, catecholamines and growth hormone. Although each can achieve the objective of restoration of blood glucose supply to the brain, the mechanism for each differs. Adrenaline (epinephrine) and glucagon are the 'rapid response' team, activating hepatic glucose formation via glycogen phosphorylation, whilst growth hormone and cortisol are the 'back up' team inducing the enzymes responsible for gluconeogenesis. Catecholamines can additionally increase circulating non-esterified fatty

acid concentrations by upregulating lipolysis so muscle uptake of glucose falls, leaving more glucose for the brain to use.

In contrast, if the counter-regulatory hormones are present in excess and there is also a relative deficiency of insulin then hyperglycaemia results; infections and trauma are common causes of increased secretion of catecholamines.

Glucose transporters

Glucose is a hydrophilic molecule that is unable directly to penetrate the lipid bilayer of cells: its uptake into cells is therefore achieved by an energy-independent process of facilitated diffusion down its concentration gradient, mediated by a family of glucose transporter proteins (GLUTs) composed of at least 12 membrane spanning helices with a larger intracellular loop connecting the sixth and seventh helices. These transporters allow uptake of glucose into cells from the interstitial fluid into which glucose diffuses from the bloodstream and, by virtue of differences in kinetics, tissue and subcellular expression profiles, and substrate specificities, mediate specific functions such as glucose sensing (GLUT2) and insulin-dependent glucose uptake (GLUT4) (Table 15.1).

The 14 facultative glucose transporters recognized to date are classified into those having high glucose affinity (class I comprising GLUTs 1–4), high fructose affinity

TABLE 15.1 Characteristics of some glucose transporters

Transporter	Tissues	Kinetics	Transport type
Facilitative glucose transporters			
GLUT1	Ubiquitous: e.g. erythrocyte, placenta, colon, kidneys	Low K_m (1–2 mmol/L)	Facilitated diffusion
GLUT2	Liver, small intestine, kidneys, β -Cells	High K_m (~25 mmol/L), high V_{max}	Facilitated diffusion; bidirectional
GLUT3	Ubiquitous: e.g. brain, placenta, kidneys	Low K_m (1–2 mmol/L), low V_{max} (6–7 mmol/L)	Facilitated diffusion
GLUT4	Skeletal muscle, adipocytes, heart	K_m 2–10 mmol/L	Facilitated diffusion, insulin responsive
GLUT5	Jejunum		Facilitated diffusion of fructose
Na⁺-glucose cotransporters (SGLTs)			
		Move glucose against concentration gradient	Active transport, symport using Na ⁺ gradient generated by K ⁺ , Na ⁺ -ATP pump
SGLT1	Intestine, renal tubules	High affinity (K_m ~0.4 mmol/L), low capacity	Intestinal glucose/galactose absorption. Some renal glucose reabsorption
SGLT2	Renal tubules	Low affinity (K_m ~2.0 mmol/L), high capacity	Most renal glucose reabsorption

K_m , substrate concentration at half the maximum velocity (V_{max}).

(class II, e.g. GLUT5) and novel transporters whose physiology is not yet fully understood (GLUTs 6–14). The different functions of the class I GLUTs are partly predictable from their differing K_m values. Glucose transporters 1, 3 and 4 have K_m values of ~2–5 mmol/L but GLUT2 has a K_m of ~25 mmol/L. This permits high rates of glucose entry into essential cells (e.g. in the central nervous system (CNS), which is relatively protected from neuroglycopenia by the low K_m of its GLUT3 transporters). Pancreatic β -Cells are able to sense increments in blood glucose over a range considerably exceeding normal values via the high K_m GLUT2 transporter protein.

Certain glucose transporters (GLUT1 and GLUT3) are present on cell surfaces at all times. In contrast, GLUT4 is stored in the cytoplasm when insulin is not present; it responds to insulin by moving from intracellular stores to cell membranes, thereby increasing total transporter number (typically 6–10-fold). When insulin concentrations decline, GLUT4 is removed from cell membranes by endocytosis and rapidly recycled back to its intracellular storage compartments. Erythrocytes are insulin non-responsive because they possess only GLUT1. Many cells can express a variety of different GLUT transporters, and the expression of GLUT receptors changes with circumstances: for example liver cells express more GLUT1 and GLUT3 during starvation. Dysfunction of the insulin-regulated GLUT4 translocation process appears to play a part in the insulin resistance syndrome, and mutations of several transporters (e.g. GLUT1, GLUT2) have been associated with inherited metabolic diseases of carbohydrate metabolism.

Tissue glucose uptake via GLUT transporters involves facilitated diffusion down a concentration gradient, the intracellular concentration of glucose being very low because of its active metabolism by phosphorylating enzymes (hexokinases and glucokinases). In resting, postabsorptive subjects, ~70% of the body's glucose metabolism occurs independently of the action of insulin. However, these insulin-independent mechanisms cannot maintain normoglycaemia for very

long without insulin orchestrating the response to food and integrating the balance between fatty acid, ketone body and glucose metabolism. Insulin-independent (as well as insulin-dependent) glucose clearance is impaired in subjects with type 2 diabetes and also in normoglycaemic subjects with a family history of diabetes. This suggests that abnormalities in insulin-independent glucose disposal manifest at a very early stage of disease development. This phenomenon of 'glucose resistance' appears to be quantitatively important: in normal subjects, as much as half an intravenous glucose load is cleared by non-insulin-mediated glucose disposal. Although polymorphisms of GLUT4 do not appear to be any more common among patients with type 2 diabetes than in the normal population, there is evidence that targeting and trafficking of this glucose transporter may be abnormal in both skeletal muscle and adipose tissue in this condition. The insulin-sensitizing agents metformin and the thiazolidinediones appear to increase cell surface expression of GLUT4, as does physical exercise.

The ability to move glucose against a concentration gradient, necessary in the special circumstances of the renal tubules and intestinal epithelium, is conferred by a family of sodium-dependent glucose cotransporters. Sodium-dependent glucose cotransporter 1 (SGLT1) is present in the gut, where it is responsible for absorption of glucose and galactose from the diet. Mutations of SGLT1 are associated with the glucose–galactose malabsorption syndrome that may cause fatal infantile diarrhoea unless these sugars are removed from the diet. Sodium-dependent cotransporter 2 (SGLT2) is a low affinity high capacity transporter that is present in the S1 segment of the proximal renal tubules and is responsible for 90% of renal glucose reabsorption, the remaining glucose being reabsorbed by the high affinity but low capacity SGLT1 transporter found in the S3 segment; reduced function mutations of SGLT2 are associated with renal glycosuria. Drugs that block SGLT2 (thus increasing urinary glucose loss) have recently been licensed in the UK as glucose-lowering agents.

The fate of intracellular glucose and its role in diabetes

Some of the principal intracellular metabolic pathways for glucose, together with their rate-limiting enzymes, are schematically illustrated in Figure 15.2, with the pathways labelled a–k.

In order to trap glucose within cells (since all the GLUTs are potentially bidirectional), glucose is phosphorylated on entry by a family of hexokinases (a). Hexokinase types I–III are widely expressed and have low K_m , but hexokinase type IV (also called glucokinase, and predominantly expressed in liver and the β -Cells of the islets of Langerhans) has a much higher K_m (up to 15 mmol/L), permitting it to function as a glucose sensor over (and beyond) the physiological range of blood glucose. Since glucokinase is also the rate-limiting step in glucose metabolism, it thus becomes the ultimate determinant of the rate of insulin secretion from β -Cells.

Loss of function mutations of glucokinase are responsible for one form of maturity onset diabetes of the young (MODY 2, which represents approximately 30% of all cases of MODY). Patients present with mildly elevated blood glucose concentrations owing to apparent elevation of the set point for blood glucose concentrations. For reasons that are not totally clear, patients with this condition appear to be at very low risk for the development of diabetic complications and are rarely symptomatic. Although a good response is typically obtained with augmentation of the insulin response by sulfonylurea therapy, this is often not necessary. Although glucokinase might be a potential candidate gene in type 2 diabetes (and a treatment target, given that activating mutations can cause hypoglycaemia in man), mutations appear to be no more common among people with type 2 diabetes than in the general population.

Dephosphorylation of glucose (the reverse reaction) is catalysed by glucose 6-phosphatase (b). This process is necessary for the export of glucose (from gluconeogenesis) by hepatic and renal cells in hypo-insulinaemic situations. Deficiency of this enzyme (or of the associated glucose 6-phosphate transport protein) is the cause of glycogen storage disease type 1 (von Gierke disease), and its overactivity is a feature of the increased and relatively insulin-insensitive hepatic glucose production typical of type 2 diabetes. The insulin sensitizing agents metformin and the thiazolidinediones appear to reduce the activity of glucose 6-phosphatase, although it is not clear whether these are direct effects or are mediated through some other upstream action.

Having entered the cell and undergone phosphorylation, glucose undergoes one of four distinct metabolic fates. It may be:

- metabolized aerobically to yield energy
- metabolized anaerobically to yield energy
- used to synthesize other molecules (glycerol and the carbon skeletons of certain non-essential amino acids)
- stored as glycogen.

The dominant flux along each pathway will depend on the tissue (e.g. gluconeogenesis can take place in the liver and kidneys, glycogen storage in liver and muscle, glycerol generation in adipose tissue) and on the prevailing metabolic circumstances such as cellular ATP requirements (requiring glycolysis), 'housekeeping' requirements (e.g. plasma membrane stability in erythrocytes), oxygen availability (determining whether glycolysis is anaerobic (yielding lactate) or aerobic (yielding pyruvate, which can be converted to acetyl-CoA and enters the tricarboxylic acid cycle) and alternative substrate availability (e.g. fatty acids or ketone bodies in the fasting state). Some of these pathways, and in particular the rate of glucose oxidation,

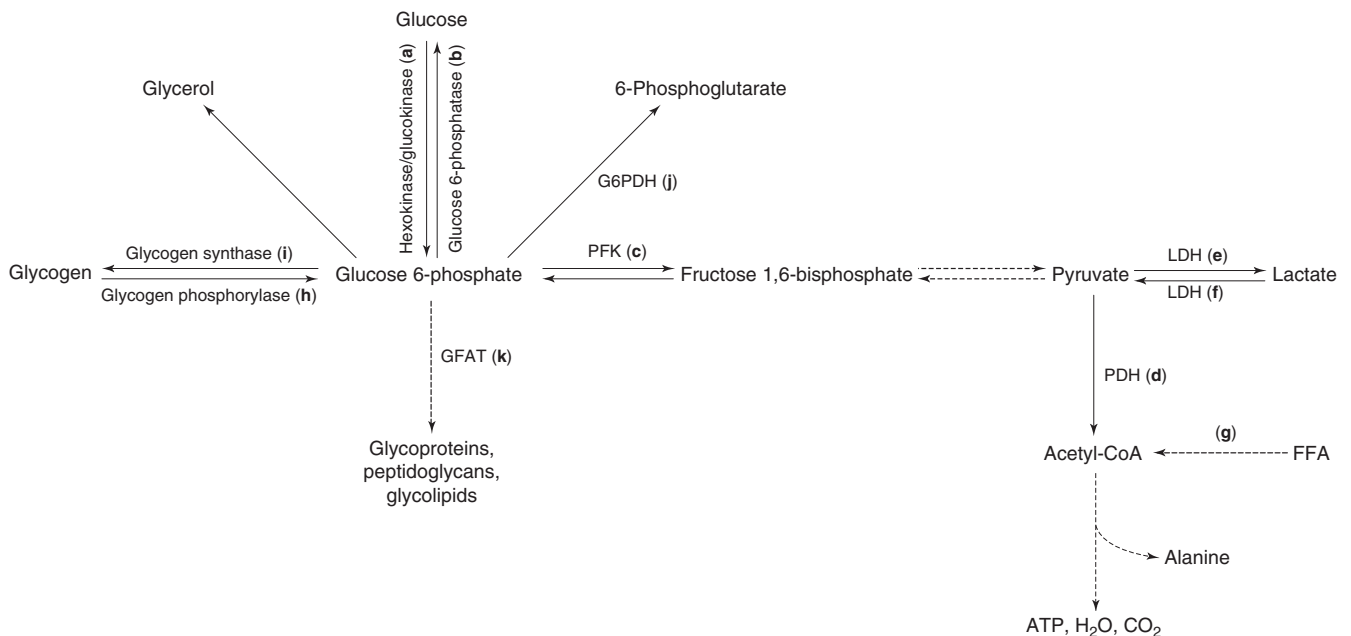


FIGURE 15.2 ■ Intracellular metabolic pathways for glucose. FFA, free fatty acids; G6PDH, glucose 6-phosphate dehydrogenase; GFAT, glutamine:fructose 6-phosphate amidotransferase; LDH, lactate dehydrogenase; PDH, pyruvate dehydrogenase. Broken lines indicate pathways from which intermediates have been omitted for clarity. Lower case letters refer to reactions, which are discussed in the text.

have been shown to be abnormal in subjects with type 2 diabetes and also in their non-diabetic relatives.

Glucose oxidation proceeds via the action of the rate-limiting enzyme phosphofructokinase (c) through various intermediates to pyruvate. Pyruvate may cross the mitochondrial membrane to enter the tricarboxylic acid cycle via the mitochondrial pyruvate dehydrogenase (PDH) enzyme complex (d) (releasing a net 38 molecules of ATP per molecule of glucose) or may remain in the cytosol and be converted anaerobically to lactate via the bidirectional enzyme lactate dehydrogenase (LDH) (e) (yielding only a net two molecules of ATP per molecule of glucose). Pyruvate dehydrogenase in muscle is stimulated by exercise and in most tissues by insulin, but this stimulation is reduced in diabetes. Pyruvate dehydrogenase is inactivated by phosphorylation, catalysed by PDH kinase (which is over expressed in insulin resistance). As a result, glucose metabolism is shifted away from the tricarboxylic acid cycle. This effect may be prevented by activation of peroxisome proliferation activating receptor α (PPAR α , see p. 720), one of a family of lipid-activated transcription factors. Mitochondrial diseases have been reported to cause insulin resistance or diabetes (e.g. diabetes insipidus, diabetes mellitus, optic atrophy and deafness (DIDMOAD) syndrome, Friedreich ataxia and HIV-associated lipodystrophy); indices of mitochondrial activity (e.g. mitochondrial number and size) are generally reduced in insulin-resistant states.

Lactate, generated under anaerobic conditions (and in cells such as erythrocytes, which lack the requisite enzymes) can be used as a substrate for gluconeogenesis either locally or after export to the liver (the Cori cycle), or may enter the tricarboxylic acid cycle after reconversion to pyruvate (f).

It has been postulated that an increased availability of substrates such as FFAs or ketone bodies, entering the tricarboxylic acid cycle via acetyl-CoA (g), may lead to excess generation of citrate, with the consequent inhibition of phosphofructokinase (PFK), thus shunting glucose down alternative, non-oxidative pathways (the glucose-FAA or Randle cycle). The consequent reduction in the entry of glucose into the tricarboxylic acid cycle could contribute to diminished glucose oxidation and thus to diabetes.

Glycogenolysis (h) is initiated by glycogen phosphorylase, the activity of which is stimulated by glucagon (in liver) and catecholamines (in muscle). In muscle, the glucose liberated from glycogen is used locally as an energy source: it is not exported into the circulation. Muscle glycogen storage via glycogen synthase (i) is diminished in type 2 diabetes, potentially contributing to reduced glucose disposal.

Intracellular glucose can also be metabolized through the pentose phosphate shunt (hexose monophosphate shunt) via glucose 6-phosphate dehydrogenase (j). Deficiency of this enzyme causes an X-linked form of haemolytic anaemia, owing to the fact that this pathway is the only route of glucose metabolism in red blood cells; it is prevalent in people originating from the Mediterranean littoral.

The rate-limiting enzyme of the hexosamine synthetic pathway (glutamine:fructose 6-phosphate amidotransferase,

GFAT) (k) is overactive in diabetes and hyperglycaemia; this pathway leads to the glycosylation of proteins, including transcription factors, and can affect cellular sensing of glucose. The hexosamine pathway is also relevant to the development of complications of diabetes through alterations in endothelial function mediated by endothelial nitric oxide synthase (eNOS), protein kinase A and protein kinase C. Defects in glucose oxidation and glycogen storage may cause shunting of glucose down this pathway, leading to insulin resistance, reduced insulin secretion and diabetic complications.

Animal models of glucotoxicity show many of the features of type 2 diabetes, for example reduced GLUT4 translocation (reducing glucose uptake into cells), reduced glycogen synthase activity (reducing glucose incorporation into glycogen), increased hepatic glucokinase activity (increasing intracellular glucose trapping), increased hepatic glucose output and reduced β -Cell glucokinase activity (reducing insulin secretion). Insulin resistance may thus be a defence against excessive intracellular accumulation of glucose, sensed by excessive shunting down the hexosamine synthetic pathway and mediated via the transcriptional regulation of key enzymes and pathways ('cellular satiety'). Reduced glucose oxidation in pancreatic β -Cells could lead to impairment of insulin secretion, and in other tissues could lead to diminished glucose disposal, both of which are hallmarks of type 2 diabetes. However, the extent to which these abnormalities actually contribute to diabetes rather than relate to specific experimental conditions (e.g. glucose and insulin concentrations), or are simply consequences of some other pathological processes, remains unclear.

Insulin

Biosynthesis

Insulin is a peptide hormone (51 amino acids arranged in two peptide chains linked by two disulphide bonds, molecular weight 5807 Da), which is secreted by the β -Cells of the islets of Langerhans in the pancreas. In the synthesis of insulin, translation of mRNA yields preproinsulin, a prohormone that undergoes post-translational modification prior to release of the mature insulin molecule. Removal of 24 amino acids from preproinsulin yields proinsulin, which consists of 86 amino acids. Proinsulin is stored in secretory granules prior to release from β -Cells by exocytosis. In healthy subjects, >90% of proinsulin is converted to mature insulin by the removal of the metabolically inert C-peptide component prior to secretion. The other products of the post-translational modification are either released when exocytosis occurs or are degraded within the secretory granules prior to release. C-peptide is co-secreted in equimolar amounts with mature insulin. In healthy subjects, only small amounts (<10% of mature insulin output) of proinsulin and partially split proinsulin are released. These ratios are characteristically disturbed in certain pathological states, including autonomous insulin secretion from an insulinoma and type 2 diabetes mellitus. Plasma concentrations of proinsulin and C-peptide are low or

undetectable in cases of surreptitious administration of exogenous insulin. Assay of these substances may, therefore, prove helpful in the differential diagnosis of hypoglycaemia in some circumstances. Proinsulin may accumulate in renal failure and its plasma concentration is elevated in familial hyperproinsulinaemia.

Substances stimulating the synthesis and storage of insulin include glucose, mannose, leucine, arginine and a variety of metabolizable sugars and sugar derivatives. Most of these also promote secretion.

Secretion and pharmacokinetics

The mechanisms by which insulin release is triggered are the focus of much research. It is apparent that there is an ATP-dependent, sulfonylurea-sensitive K^+ channel whose closure is a late event in the intracellular signalling mechanism within the β -Cell; closure triggers calcium influx and exocytosis. It is not clear how this K^+ channel is activated, although a wide range of secretagogues can stimulate activation of this final common pathway. The most important of these is hyperglycaemia, although mannose, lactate, arginine, leucine and other amino acids, glucagon, glucagon-like peptide 1 (GLP-1), glucose-dependent insulinotropic peptide (GIP, known formerly as gastric inhibitory peptide), cholecystokinin, vasoactive intestinal peptide (VIP), sulfonylureas and parasympathetic cholinergic (muscarinic) nerve activity also stimulate insulin release; many of these secretagogues have synergistic effects. Both neural sympathetic tone and circulating catecholamines inhibit insulin secretion.

Because of cephalic and gastric influences, oral glucose is a more potent stimulus to insulin secretion than an equivalent amount of intravenous glucose. This difference is known as the 'incretin' effect and is predominantly mediated by gut-derived hormones such as the N-terminally truncated GLP-1 (7–36) amide, and GIP. Recently, drugs have become available that augment the release of insulin via the sulfonylurea-sensitive K^+ channels of β -Cells (meglitinide analogues) and GLP-1 receptors (GLP-1 analogues), both acting in a glucose-dependent manner. These mechanisms afford the clinically desirable possibility of augmenting insulin release during conditions of hyperglycaemia, while at the same time reducing the risk of hypoglycaemia. They are discussed further in Chapter 16.

In healthy adults, insulin is secreted in pulses with a pulse periodicity of 11–15 min. Stimuli of insulin secretion increase the frequency and amplitude of these pulses. Approximately 30–40 U (210–280 nmol) of insulin are secreted per 24 h in healthy subjects of normal weight. Insulin secretion is basal (0.25–1.0 U/h) until glucose concentration exceeds a threshold of ~5 mmol/L, and becomes maximal at a concentration of 15–20 mmol/L.

Insulin is secreted into the portal venous system and thus must pass through the liver before reaching the systemic circulation. Approximately half of the insulin is cleared in the 'first pass' through the liver. The liver is exposed to insulin concentrations approximately threefold higher than other tissues during endogenous insulin secretion. The first pass clearance of insulin by the liver is variable, being controlled not only by the prevailing blood glucose concentration but also by factors released from the

gut before absorption is complete, and by the integrated action on β -Cell receptors of a range of endocrine and nutritional signals. It is, therefore, not surprising that it is difficult to replicate physiological patterns of insulin concentrations by giving exogenous insulin subcutaneously.

Autocrine and paracrine regulation of insulin secretion by pancreatic and gut hormones (which may reach very high concentrations within islets) is not well understood. Increased secretion of insulin involves recruitment of more β -Cells to the secreting mode.

Fasting peripheral insulin concentrations vary between 20 and 140 pmol/L (~2.9–20 mU/L), as measured by immunoassays in healthy subjects, the higher values being associated with increasing age and obesity. After a typical mixed meal (700–800 kcal), peak plasma insulin concentrations reach ~350–580 pmol/L (50–84 mU/L) in lean young adults. The half-life of insulin injected into a peripheral vein is 2–6 min, with most being cleared by the liver and smaller amounts being cleared by other tissues that have insulin receptors, such as skeletal muscle, although there is also non-receptor mediated clearance by a variety of tissue proteases.

Abnormalities of the synthesis and secretion of insulin

There are a number of recognized genetic abnormalities of insulin structure involving mutations of the insulin gene and hence altered amino acid sequences. Some of these are listed in Table 15.2. Each of the proinsulin and insulin variants mentioned has reduced biological activity in relation to insulin. This causes a propensity to diabetes, although individuals who can sustain a compensatory hypersecretion may not develop the condition. For example, in familial hyperproinsulinaemia, hypersecretion is usually sufficient to prevent diabetes, whereas in insulin Wakayama, diabetes is usual. There are also recognized associations between some polymorphisms affecting the insulin secretory mechanism (e.g. calpain 10, a molecule that promotes the fusion of secretory granules with cell membranes) and diabetes.

More common abnormalities of insulin secretion involve loss of the normal pulsatility of insulin release, an early feature of both obesity and type 2 diabetes. The progressive loss of insulin secretory capacity in diabetes is discussed below.

Actions of insulin

Insulin has widespread actions, some of which are listed in Table 15.3. It is the dominant hormone regulating blood glucose concentration. It should be noted that while the mechanisms of its glucoregulatory action have been the subject of extensive research, much less is known about its other actions in health or in insulin-resistant states, although vascular effects (vascular smooth muscle proliferation, vasodilatation), CNS effects (appetite, learning, memory) and effects in relation to growth, differentiation and apoptosis have all been proposed.

At present, only two receptors that mediate the actions of insulin have been identified: the insulin receptor and the IGF (insulin-like growth factor) receptor. It is apparent,

TABLE 15.2 Some genetically determined abnormal insulins

Condition	Abnormality	Consequences
Familial hyperproinsulinaemia	Failure to cleave C-peptide from proinsulin	True insulin concentrations and glucose tolerance normal
Insulin Chicago	Leu for Phe substitution at position 25 of β chain	Reduced receptor binding
Insulin Los Angeles	Ser for Phe substitution at position 24 of β chain	Reduced receptor binding
Insulin Wakayama	Leu for Val substitution at position 3 of α chain	Reduced receptor binding
Proinsulins Boston and Tokyo	His for Arg substitution at position 65	Inability to cleave proinsulin to insulin
Proinsulin Providence	Asp for His substitution at position 10 of β chain	Inability to cleave proinsulin to insulin
Proinsulin Kyoto	Leu for Arg substitution at position 65	Inability to cleave proinsulin to insulin

TABLE 15.3 The principal actions of insulin

Actions	Mechanisms
Liver	
Inhibition of hepatic glucose output	Limitation of substrate supply Inhibition of glycogenolysis Inhibition of gluconeogenesis
Stimulation of hepatic glycogen storage	Stimulation of glycogen synthase
Stimulation of hepatic glycolysis for intermediary metabolism	Stimulation of phosphofruktokinase
Stimulation of hepatic lipogenesis	Stimulation of pyruvate dehydrogenase
Stimulation of hepatic glucose oxidation	Stimulation of pyruvate dehydrogenase
Skeletal muscle	
Stimulation of glucose transport	Activation of glucose transporter (GLUT4)
Stimulation of muscle glycogen synthesis	Stimulation of glycogen synthase
Stimulation of muscle glycolysis	Stimulation of phosphofruktokinase
Adipose tissue	
Inhibition of lipolysis (stored lipid)	Inhibition of hormone-sensitive lipase
Promotion of re-esterification	? Increased supply of glycerol 3-phosphate
Stimulation of lipolysis (circulating lipid)	Stimulation of lipoprotein lipase
Increased glucose uptake	Several (? as for muscle/liver)
Central nervous system	
Satiety	Uncertain
Changes in sympathetic tone	Uncertain
Postprandial thermogenesis	Uncertain
Other	
Promotes DNA synthesis	Uncertain
Promotes RNA synthesis	Various
Stimulation of amino acid uptake	Uncertain
Na ⁺ ,K ⁺ -ATPase stimulation	? Increase in intracellular energy availability
Na ⁺ /H ⁺ antiport activation	Uncertain
Sodium retention	Probably several mechanisms

however, that there are individual dose–response curves for the different actions of insulin in different tissues. For example, the median effective dose (ED₅₀) for insulin's antilipolytic action on adipose tissue is <140 pmol/L (20 mU/L) (and may be <70 pmol/L), while those for inhibition of hepatic glucose output and stimulation of glucose uptake into skeletal muscle are 210–350 pmol/L (30–50 mU/L) and 350–490 pmol/L (50–70 mU/L), respectively. A doubling of insulin concentration inhibits hepatic glucose output by ~80% and stimulates peripheral glucose utilization by ~20%. In patients with type 2 diabetes, these differential effects on lipolysis, hepatic glucose output and glucose uptake, coupled with ongoing (if reduced) insulin secretion, are probably responsible

for the fact that most individuals are not at risk of developing ketoacidosis (at least for many years), despite the clear defect in glucoregulation that is evident.

The different actions of insulin have different time courses, with the glucoregulatory and antilipolytic occurring within a few minutes, and growth regulation and actions dependent upon synthesis of new proteins occurring over periods of hours or days. Intravenous injection of insulin typically has little effect on blood glucose for 5–10 min, the maximal hypoglycaemic action occurring after 5–15 min. Insulin stimulation of skeletal muscle glucose uptake declines with a half-life of 10–20 min after the insulinaemic stimulus has been removed.

Proinsulin and partially split proinsulins have metabolic activities generally similar to those of insulin, although their plasma half-lives are 3–5 times longer and their biological potencies are only 8–15% that of insulin. It has been suggested that the hepatic activity of proinsulin may be relatively more potent than its effect on peripheral glucose uptake.

The insulin receptor

The main glucoregulatory effects of insulin are mediated by a transmembrane receptor found on insulin-sensitive cells. This receptor is a glycoprotein, with total molecular weight of 350 000 Da, comprising four peptide chains (two α - and two β -subunits), linked by disulphide bridges (Fig. 15.3). Two isoforms of the receptor (IR-A and IR-D) are formed by alternative splicing. The gene for, and amino acid structure of, the insulin receptor have been characterized and show homology with those of the IGF-1 receptor (see below). Within the intracellular domain of the β -subunit is a tyrosine kinase capability, which is activated when insulin binds with the extracellular domain of the α -subunits. The tyrosine kinase promotes autophosphorylation of the receptor followed by activation of threonine and serine kinases.

The insulin receptor gene is located on the short arm of chromosome 19 (19p 13.2). Rare gene mutations have been described, for example leprechaunism and Rabson–Mendenhall syndrome, which result in severe glucose intolerance with resistance to exogenous insulin, and profoundly disordered growth, rather than typical insulin resistance. These conditions are usually lethal in infancy and adolescence, respectively. There are also more common, ‘milder’ polymorphisms of the insulin receptor gene. However, these appear to explain only a small proportion of the marked variance in population insulin sensitivity and they are considered to be a rare (<1%) cause of type 2 diabetes mellitus. Indeed, most mutations of the insulin gene that have been recognized are not sufficient to cause diabetes alone, although they may render it more likely to occur in the presence of other risk factors.

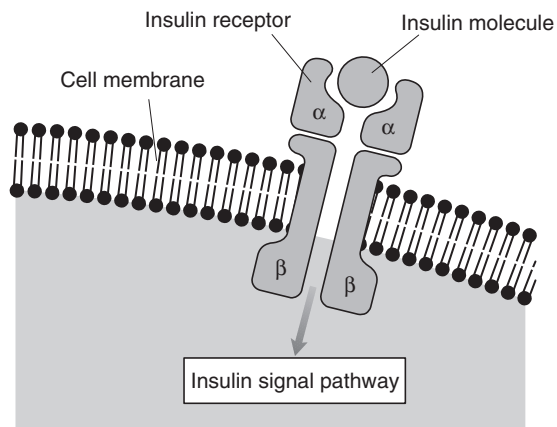


FIGURE 15.3 ■ The insulin receptor.

Second messengers mediating the effects of insulin

Insulin can have multiple actions, even on a single responsive cell, and hence there are probably several different intracellular pathways mediating its actions. Glucoregulatory and antilipolytic responses are rapid and probably mediated via serine and threonine kinases and cAMP. Stimulation of lipid and protein synthesis, inhibition of proteolysis, the nuclear transcription of RNA and the replication of DNA are slower and have different second messenger systems. Second messenger systems involving diacylglycerol, protein kinase C and glycosyl-phosphatidyl-inositol are under investigation, but abnormalities in them have yet been demonstrated to be necessary or sufficient to cause diabetes in man. As a result of these second messenger cascades, glucose transporter proteins are translocated from within cells to their surface membranes to increase glucose flux into the cytoplasm.

After activation, insulin–receptor complexes are internalized by an endocytotic process; receptors are later recycled to the cell surface. Internalization of the insulin is important (and possibly essential) for insulin signals to reach the nucleus and influence cell growth and protein synthesis. Internalization is an important route by which insulin is cleared from the circulation and degraded. The actions of insulin in stimulating DNA transcription and mRNA translation do not depend upon the plasma membrane insulin receptor kinase activity and the second messenger systems discussed above, or on the IGF receptors described below, but rather involve direct effects within the nucleus and on ribosomes.

Insulin-like growth factors and their receptors

In addition to its acute effects on glucose uptake and release and on lipid metabolism, insulin has growth promoting activity in a variety of tissue culture models. At least two protein hormones, insulin-like growth factors 1 and 2 (IGF-1 and IGF-2), have actions that partly resemble these actions of insulin. The amino acid sequences of these proteins and the base sequences of their coding DNA are known and show some homology with those of insulin. Insulin-like growth factors are weak agonists for the insulin receptor and hence have weak glucoregulatory and antilipolytic effects. In addition, they have growth-promoting effects mediated by two IGF receptors. Insulin itself binds only weakly to IGF receptors.

The physiological role of IGFs is not yet fully established. It has been suggested that they act as somatomedins (they were previously called somatomedins C and A, terms that are no longer recommended), in that they are induced by growth hormone and mediate its growth-promoting effects in children. In adults, they are believed to promote the growth of continually dividing cells such as osteoblasts, chondrocytes, fibroblasts and erythroid cells.

Adult growth hormone deficiency is manifest by low overall quality of life scores, reduction in lean body mass, centripetal fat distribution, loss of bone mineral,

abnormal lipid profile, insulin resistance and changes in the secretion and metabolism of other circulating hormones, for example thyroxine (see p. 367). The effects of growth hormone deficiency on insulin sensitivity are complex. While glucose intolerance is a common feature of acromegaly, in growth hormone deficiency states, adverse changes consequent upon an increase in body fat composition tend to outweigh the lack of a counter-regulatory hormone to produce insulin resistance overall. Replacement therapy in adult growth hormone deficiency produces somewhat unpredictable effects, although a period of increased insulin resistance may herald a modest longer-term improvement.

Tumour-related hypoglycaemia may be due to excess production of IGFs, specifically 'big' IGF-2. Local and systemic concentrations of IGFs are modulated by a range of binding proteins. While total IGF-2 may not be increased in tumour-related hypoglycaemia, the IGF-2 is present in an immature form which, because of impaired ability to form the normal ternary 150 kD complex with IGF binding protein-3 (IGFBP-3) and the acid-labile subunit (ALS), circulates with greater than normal bio-availability. This effect has been described in a number of tumours, especially those of mesenchymal and epithelial origin. Further discussion of this topic is presented in Chapter 17.

It was first observed many years ago that proliferative diabetic retinopathy could regress following hypophysectomy in both animals and humans. This effect appears to be mediated via reduction in IGF-1, owing to the growth hormone deficiency seen after pituitary ablation. It has been hypothesized that the metabolic derangements of diabetes reduce hepatic IGF-1 generation in response to growth hormone (i.e. induce growth hormone resistance), which is then hypersecreted (owing to a lack of negative feedback from IGF-1), reaching concentrations that are able to stimulate IGF-1 generation in non-hepatic tissues. However, studies of pegvisomant, a growth hormone analogue that antagonizes IGF-1 generation in response to growth hormone, have not shown benefit in diabetic retinopathy.

CLASSIFICATION AND DIAGNOSIS OF DIABETES MELLITUS

Introduction

Diabetes is the most common metabolic disorder, with a cumulative incidence of 5–10% in people in the developed world aged >40 years, in whom >90% of cases are due to type 2 diabetes. Population screening programmes typically reveal that up to half of the subjects found to have type 2 diabetes had previously been undiagnosed. While the burden of type 2 diabetes is increasing exponentially with the epidemic of obesity in many parts of world, the incidence of type 1 diabetes has also been increasing for many years for reasons that are much less apparent. Nevertheless, the prevalence of type 2 diabetes in children is approaching that of type 1 diabetes, having been only ~2–4% of all childhood diabetes prior

to 1994, and is predicted to outstrip type 1 diabetes by 2025 on current trends. A high proportion of these children, typically presenting around the time of puberty, require insulin from the time of diagnosis. Whether this is due to the accelerator hypothesis ('double diabetes effect'), whereby an individual's risk of contracting type 1 diabetes is increased by the prior existence of type 2 diabetes or insulin resistance, is not clear.

Although insulin has many actions, as described above, diabetes is defined only in terms of elevated blood glucose concentrations. Since blood glucose is a continuous variable, the cut-off points for diagnosis are necessarily somewhat arbitrary.

Recent changes to the diagnostic criteria for diabetes and glucose intolerance reflect the increased cardiovascular risk evident at even moderate levels of fasting hyperglycaemia (~6.0 mmol/L in some studies). However, the blood glucose threshold for this effect is almost certainly lower than that for the predominantly microvascular complications unique to diabetes mellitus. Some people so diagnosed may not, therefore, necessarily be at risk of developing the unique set of complications (nephropathy, retinopathy, neuropathy) that have traditionally characterized the disease and determined its management.

Definitions

The diagnostic criteria for diabetes mellitus set out by The World Health Organization (WHO) have been adopted by both the American Diabetes Association (ADA) and Diabetes UK, although they differ slightly for glucose intolerance. The WHO introduced diagnostic criteria in 1979, but introduced revised criteria in 2000 to reflect better understanding of 'milder' glucose intolerance. The use of glycated haemoglobin (HbA_{1c}) for diagnosis was introduced in 2011. The WHO glucose criteria for diagnosis are shown in [Box 15.2](#) and [Table 15.4](#) and its recommendations regarding the use of HbA_{1c} are shown in [Box 15.3](#). World Health Organization criteria consider fasting and 120 min values in the oral glucose tolerance test (OGTT). The reproducibility of the OGTT leaves much to be desired (the CV of 120 min plasma glucose concentrations is reported to be up to 50%). If a subject fulfils the WHO criteria for diabetes, subsequent improvement of glucose tolerance may occur (e.g. as a result of weight loss or spontaneously), but such individuals are considered to have a lifelong tendency to diabetes.

[Table 15.4](#) indicates the criteria for the diagnosis of impaired glucose tolerance (IGT) and impaired fasting glycaemia (IFG), which are metabolic states intermediate between normal glucose tolerance and diabetes. For epidemiological or population screening purposes, the fasting or 2 h value after 75 g oral glucose may be used alone. For clinical purposes, the diagnosis of diabetes should always be confirmed by repeating the test on another day, unless there is unequivocal hyperglycaemia with acute metabolic decompensation or obvious symptoms. Glucose concentrations should not be determined on serum or plasma unless red cells have been immediately

BOX 15.2 World Health Organization recommendations for glucose-based diagnosis of diabetes mellitus

Criteria for the diagnosis of diabetes mellitus

1. Symptoms of diabetes plus casual plasma glucose concentration ≥ 11.1 mmol/L. Casual is defined as any time of day without regard to time since last meal. The classic symptoms of diabetes include polyuria, polydipsia and unexplained weight loss
- or*
2. Fasting plasma glucose ≥ 7.0 mmol/L. Fasting is defined as no caloric intake for at least 8 h
- or*
3. Two-hour postload glucose ≥ 11.1 mmol/L during an oral glucose tolerance test (OGTT). The test should be performed as described by WHO, using a glucose load containing the equivalent of 75 g anhydrous glucose dissolved in water.

In the absence of unequivocal hyperglycaemia or classic symptoms, these criteria should be confirmed by repeat testing on a different day.

The third measure (OGTT) is not recommended for routine clinical use.

removed, or glycolysis will result in an unpredictable underestimation of the true concentrations. It should be stressed that glucose preservatives such as fluoride do not totally prevent glycolysis. If whole blood is used, the sample should be kept at 0–4°C, and assayed as soon as possible.

Subjects with IGT and IFG are at high risk of progression to diabetes; IGT and IFG should therefore be considered as risk categories for future diabetes and/or cardiovascular disease. In certain cases, glycated haemoglobin (HbA_{1c}), the concentration of which reflects glycaemia over several weeks, gives sensitivity and specificity for diagnosis almost equal to that of glucose measurements, and a cut-off value of 48 mmol/mol (6.5%) is now accepted as diagnostic, although it is important

to note that a value below this level does not necessarily exclude diabetes. Several classifications of diabetes have been proposed. The most widely used is that proposed by the WHO and adopted by the ADA shown in Box 15.4. However, it should be recognized that unanimity in nomenclature has yet to be achieved, especially in the areas of gestational diabetes, diabetes related to pancreatitis and tropical/malnutrition-related diabetes.

Type 1 diabetes mellitus

Introduction

Approximately 5–10% of all patients with diabetes have type 1 diabetes mellitus: this is characterized by severe insulin deficiency because of β -Cell destruction. The degree of insulin deficiency is so severe that patients require exogenous insulin therapy to avoid the rapid decline into cachexia, dehydration, ketoacidosis and death that was the inevitable consequence of this disease before the discovery of insulin by Banting and Best in 1921. Dietary measures cannot prevent this problem, although they may delay its onset, and survival for several years in the pre-insulin era with harshly restrictive hypoketotic diets was not unknown. Withdrawal of exogenous insulin therapy in patients with type 1 diabetes usually results in ketoacidosis within 48 h, and this may supervene in less than 12 h when insulin has previously been administered intravenously (and therefore cleared within minutes with no depot effect) and in conditions of increased insulin requirement or counter-regulatory response such as systemic sepsis.

The severity of type 1 diabetes is such that most patients come to medical attention within weeks or, at most, months of onset, whereas the onset of type 2 diabetes is usually more insidious, so it may be present many years before diagnosis. Although the condition (including latent autoimmune diabetes of adults, LADA, see below) can present at any age, most patients present in childhood and the diagnosis becomes rare after 30 years of age.

TABLE 15.4 Glucose concentrations for the diagnosis of diabetes mellitus and other categories of hyperglycaemia in accordance with WHO guidelines (1999 and 2006)

	Glucose concentration (mmol/L)			
	Whole blood		Plasma	
	Venous	Capillary	Venous	Capillary
Diabetes mellitus				
Fasting	≥ 6.1	≥ 6.1	≥ 7.0	≥ 7.0
<i>or</i>				
2 h post-glucose load	≥ 10.0	≥ 11.1	≥ 11.1	≥ 12.2
<i>or both</i>				
Impaired glucose tolerance (IGT)				
Fasting (if measured)	< 6.1	< 6.1	< 7.0	< 7.0
<i>and</i>				
2 h post-glucose load	6.7–9.9	7.8–11.0	7.8–11.0	8.9–12.1
Impaired fasting glycaemia (IFG)				
Fasting	5.6–6.0	5.6–6.0	6.1–6.9	6.1–6.9
<i>and</i>				
2 h post-glucose load (if measured)	< 6.7	< 7.8	< 7.8	< 8.9

BOX 15.3 Diagnosis of diabetes mellitus using HbA_{1c} in accordance with World Health Organization recommendations**Criteria for the diagnosis of diabetes mellitus**

- HbA_{1c} can be used as a diagnostic test for diabetes providing that:
 - stringent quality assurance tests are in place
 - assays are standardized to criteria aligned to the international reference values
 - there are no conditions present which preclude its accurate measurement
- An HbA_{1c} of 48 mmol/mol (6.5%) is recommended as the cut-off point for diagnosing diabetes. A value <48 mmol/mol (6.5%) does not exclude diabetes diagnosed using glucose tests
- In patients with an HbA_{1c} of ≥48 mmol/mol but no symptoms of diabetes, repeat the HbA_{1c} measurement on a further sample. If the second result is <48 mmol/mol (6.5%), treat as high diabetes risk and repeat the test in six months, or sooner if symptoms develop.

Situations where HbA_{1c} is not appropriate for diagnosis of diabetes

- All children and young people
- Patients of any age suspected of having type 1 diabetes
- Patients with symptoms of diabetes for less than two months
- Patients at high diabetes risk who are acutely ill (e.g. those requiring hospital admission)
- Patients taking medication that may cause rapid glucose rise, e.g. steroids, antipsychotics

- Patients with acute pancreatic damage, including pancreatic surgery
- Pregnancy
- Haemoglobinopathies
- Anaemia (haemolytic and iron deficiency)
- Renal failure
- HIV infection.

Patients whose HbA_{1c} is <48 mmol/mol (6.5%)

- Patients may still fulfil the WHO glucose criteria for the diagnosis of diabetes
- Use WHO glucose testing in patients with symptoms of diabetes or clinically at very high risk of diabetes (although the use of such glucose tests is not recommended routinely in this situation).

Patients who are clinically at high risk of developing diabetes:

- HbA_{1c} 42–47 mmol/mol (6.0–6.4%):
 - Provide intensive lifestyle advice
 - Warn patients to report symptoms of diabetes
 - Monitor HbA_{1c} annually
- HbA_{1c} <42 mmol/mol (6.0%):
 - Patients may still have a high diabetes risk
 - Review the patient's personal risk and treat as 'high diabetes risk', as clinically indicated.

However, the distinction between types 1 and 2 diabetes is not always clear-cut and it is not always possible to categorize patients on clinical grounds alone, even after a detailed history taking into account such factors as the duration of symptoms and degree of weight loss, so that in some patients additional investigations are required to establish the correct diagnosis. Latent autoimmune diabetes of adults is a rare subtype of type 1 diabetes. Affected individuals tend to develop diabetes more insidiously and appear to respond, at least initially, to treatments other than insulin such as sulfonylureas. Some patients with autoimmune type 1 diabetes, even though they may have presented with ketoacidosis, retain sufficient islet cell reserve for some months after the diagnosis to remain ketosis free on little or even no insulin (the so-called 'honeymoon period'). Conversely, some patients with type 2 diabetes may become ketosis prone as β -Cell failure progresses from being relative to being absolute in the face of islet cell exhaustion or destruction, particularly in conditions of increased insulin requirement such as systemic sepsis. The 'accelerator hypothesis' of type 2 diabetes predisposing to type 1 diabetes via islet cell stress, and the high prevalence of the co-aggregation of features of both type 1 and type 2 diabetes is discussed above in relation to childhood onset diabetes.

Where confirmation of type 1 (autoimmune) diabetes is required, quantification of anti-glutamic acid decarboxylase (GAD) antibodies and/or islet cell antibodies (ICA) may be helpful. However, these antibodies may be detectable in healthy individuals (~2%), and the likelihood of a positive result in true autoimmune diabetes

declines from >85% at disease onset to negligible rates after 10–15 years of disease. Although normal basal or glucagon-stimulated concentrations of C-peptide (the normal response is a doubling of C-peptide concentration 6–20 min after glucagon 1 mg i.m.) rule out the existence of severe absolute insulopaenia at the time of the test, negative results are potentially harder to interpret, as endogenous insulin secretion may have been suppressed by long-term treatment with exogenous insulin.

Individuals who do not have diabetes but who have detectable anti-GAD or ICAs may be at increased future risk of developing type 1 diabetes (and other autoimmune conditions). Several studies have attempted to determine whether intervention with immunosuppressive therapy reduces this risk in those at highest risk (usually first-degree relatives of people with type 1 diabetes, with high-risk genotype and positive antibodies), or helps to preserve any residual β -Cell function at the time of diagnosis. Thus far, the side-effects of the drugs used would appear to outweigh their benefit, although some delay in progression to complete insulin deficiency has been demonstrated, at least in some of the studies.

Aetiology: genetic susceptibility and possible environmental cofactors

Type 1 diabetes typically results from an autoimmune attack upon pancreatic β -Cells. It is not understood what factors may trigger the attack. Although viral infection can provoke the attack (as seen clearly with congenital rubella), for most patients it is not clear what factor(s)

BOX 15.4 Aetiological classification of diabetes mellitus**I. Type 1 diabetes (β -Cell destruction, usually leading to absolute insulin deficiency)**

- A. Immune mediated
- B. Idiopathic

II. Type 2 diabetes (may range from predominantly insulin resistance with relative insulin deficiency to a predominantly secretory defect with insulin resistance)**III. Other specific types**

- A. Genetic defects of β -Cell function
 1. Chromosome 12, hepatocyte nuclear factor (HNF-1A) (MODY 3)
 2. Chromosome 7, glucokinase (MODY 2)
 3. Chromosome 20, HNF-4A (MODY 1)
 4. Chromosome 13, insulin promoter factor-1 (IPF-1; MODY 4)
 5. Chromosome 17, HNF-1B (MODY 5)
 6. Chromosome 2, NeuroD1 (MODY 6)
 7. Mitochondrial DNA
 8. Others
- B. Genetic defects in insulin action
 1. Type A insulin resistance
 2. Leprechaunism
 3. Rabson–Mendenhall syndrome
 4. Lipotrophic diabetes
 5. Others
- C. Diseases of the exocrine pancreas*
 1. Pancreatitis
 2. Trauma/pancreatectomy
 3. Neoplasia
 4. Cystic fibrosis
 5. Haemochromatosis
 6. Fibrocalcific pancreatopathy
 7. Others
- D. Endocrinopathies*
 1. Acromegaly
 2. Cushing syndrome
 3. Glucagonoma
 4. Pheochromocytoma
 5. Hyperthyroidism
 6. Somatostatinoma

7. Aldosteronoma
8. Others
- E. Drug- or chemical-induced*
 1. Vacor
 2. Pentamidine
 3. Nicotinic acid
 4. Glucocorticoids
 5. Thyroid hormones
 6. Diazoxide
 7. β -Adrenergic agonists
 8. Thiazides
 9. Dilantin
 10. α -Interferon
 11. Others
- F. Infections*
 1. Congenital rubella
 2. Cytomegalovirus
 3. Others
- G. Uncommon forms of immune-mediated diabetes*
 1. 'Stiff-man' syndrome
 2. Anti-insulin receptor antibodies
 3. Others
- H. Other genetic syndromes sometimes associated with diabetes
 1. Down syndrome
 2. Klinefelter syndrome
 3. Turner syndrome
 4. Wolfram syndrome
 5. Friedreich ataxia
 6. Huntington chorea
 7. Laurence–Moon–Biedl syndrome
 8. Myotonic dystrophy
 9. Porphyria
 10. Prader–Willi syndrome
 11. Others

IV. Gestational diabetes mellitus (GDM)

Statistical risk classes (subjects with normal glucose tolerance but substantially increased risk of developing diabetes)

- Previous abnormality of glucose tolerance
- Potential abnormality of glucose tolerance

*Marked causes are termed 'secondary' diabetes.
(From WHO Study Group on Diabetes Mellitus.)

is/are responsible. In those patients with other autoimmune diseases (e.g. Addison disease, autoimmune thyroid disease, coeliac disease), it appears that the attack on the pancreas is part of a multiorgan or polyglandular process. It is marked by systemic indices of the autoimmune process such as circulating islet cell autoantibodies (which, when in high titre, may be complement fixing) and changes in circulating T and B lymphocyte subsets. The process of β -Cell destruction usually takes many months and occurs in cycles of deterioration and remission. The acute presentation of type 1 diabetes mellitus usually occurs during an acute deterioration of β -Cell function, the disorder frequently being unmasked by an intercurrent illness. After the patient recovers from the intercurrent illness, adopts a suitable diet and is treated with exogenous insulin, there is often a temporary period of improved β -Cell function, known clinically as the

'honeymoon period', during which glycaemic control may be relatively easily maintained. This honeymoon period typically lasts six, but occasionally up to 24 months, and β -Cell function may be misleadingly good during this period.

Subsequently, insulin and C-peptide secretion are almost completely lost (maximal C-peptide response following a glucose load $<0.5 \mu\text{g/L}$); however, the degree of any residual β -Cell function may determine the ease with which good glycaemic control can be achieved with exogenous insulin therapy.

Although β -Cell destruction is the cause of type 1 diabetes mellitus, secondary metabolic defects occur, including resistance to exogenous insulin, although this is more commonly seen in type 2 diabetes mellitus. There are a number of mechanisms for this, including the development of anti-insulin antibodies (relatively rare with synthetic human insulins) and abnormal body composition

due in part to the peripheral administration of insulin, or because the individual is predisposed to insulin resistance by the genetic and morphological factors that also operate in type 2 diabetes mellitus. These morphological factors are becoming increasingly common with the worldwide increase in obesity, which affects individuals with type 1 diabetes as well as those without.

Certain genetic markers are associated with a high prevalence of type 1 diabetes. Most of these are found on chromosome 6, in genes related to histocompatibility linked antigens (HLAs), which together probably account for about half the genetic risk, although recent genome-wide association studies have identified several other markers of risk. Some of the associations are listed in Table 15.5. Most type 1 diabetes associated with HLA-DR4 presents in childhood while that associated with HLA-DR3 has a more variable age of onset.

Type 2 diabetes mellitus

Introduction: the heterogeneity of type 2 diabetes

Type 2 diabetes mellitus is the commonest form of diabetes mellitus worldwide. It is often considered a diagnosis of exclusion, that is, patients are assumed to have type 2 diabetes because they do not demonstrate the typical features of type 1.

Type 2 diabetes is probably not a single condition. In all patients with the condition, there is both insulin resistance and relative insulin deficiency. Some patients with late onset diabetes initially presumed to be type 2 will turn out to have type 1 (latent autoimmune diabetes in adults, LADA). With time (over perhaps 5–15 years from diagnosis), glycaemic control in type 2 patients usually becomes more difficult to achieve; insulin deficiency becomes more apparent and a subgroup of patients becomes prone to ketosis. Data from the UK Prospective Diabetes Study (UKPDS) trial suggested that the average time to insulin use was approximately seven years from diagnosis of type 2 diabetes, and confirmed the clinical impression of a progressive rather than static disease process. The typical patient with type 2 diabetes mellitus is overweight (average BMI at presentation $>27 \text{ kg/m}^2$), with a central

distribution of obesity (most conveniently assessed by waist circumference or waist:hip ratio) conferring risk that is independent of and additional to that of elevated body mass index (BMI). The age-adjusted relative risk for diabetes begins to increase at values of BMI that are considered normal based on mortality risk (24 kg/m^2 for men, 22 kg/m^2 for women), and rises exponentially as BMI rises (Fig. 15.4). The marked increase in the prevalence of obesity is an important contributor to the increased prevalence of type 2 diabetes. Increases in abdominal fat mass, weight gain since young adulthood and a sedentary lifestyle are additional obesity-related risk factors for diabetes. In some ethnic groups (particularly those from parts of south Asia), the risk of diabetes may be higher at lower levels of obesity.

Other independent environmental risk factors include being born to a mother with gestational diabetes mellitus, being of exceptionally high birth weight and being of exceptionally low birth weight. The latter is postulated by the 'Barker hypothesis' to predispose to diabetes and obesity by, among other things, switching on 'thrifty' genes to counter the effects of intrauterine malnutrition.

Leaner patients with type 2 diabetes tend to show more severe insulin deficiency. Greater degrees of obesity are associated with more insulin resistance. A major unresolved controversy of type 2 diabetes remains whether, for the typical type seen in European individuals, the prime defect in glucose homeostasis is insulin deficiency or insulin resistance or both. Given that many individuals with severe insulin resistance do not have diabetes and that some patients with type 2 diabetes have little insulin resistance, it is probably true to say that insulin resistance is neither a necessary nor sufficient cause: rather, some degree of β -Cell dysfunction (whether as an inherited tendency or as a result of reduced β -Cell function as part of a degenerative process) is the *sine qua non* of type 2 diabetes. Such β -Cell dysfunction may take the form of a relative lack of insulin secretion or of abnormal patterns of insulin secretion. Such abnormalities have been described in patients who later developed type 2 diabetes, and include changes in the amplitude and frequency of insulin secretory pulses and the loss of first-phase insulin secretion (the initial pulse of insulin secretion seen after a meal

TABLE 15.5 Genetic associations with type 1 diabetes mellitus

Allele	Chromosome	Notes
HLA-DR3	6	Relative risk = $\times 5$
HLA-DR4	6	Relative risk = $\times 7$
HLA-B8, B15 and B18	6	Due to linkage disequilibrium with DR3 and DR4
HLA-DQR4	6	Relative risks up to $\times 90$
Insulin gene	11	
T cell receptor	?	Importance disputed
Ig heavy chain	14	Importance disputed
Kidd blood group	18	

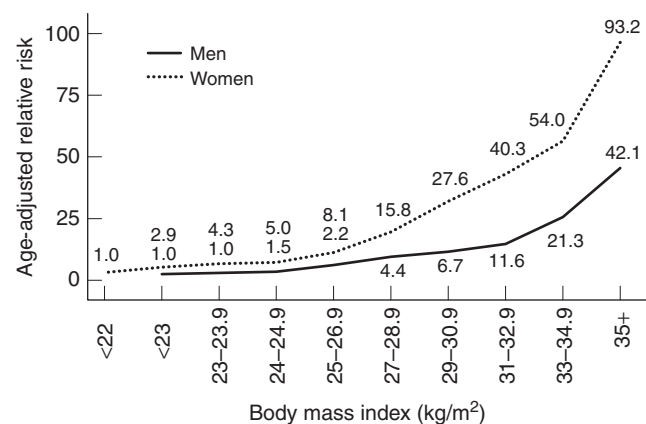


FIGURE 15.4 ■ Relationship between BMI and risk of type 2 diabetes. The risk of diabetes increases with increasing BMI values in men and women.

or glucose tolerance test), with prolongation and augmentation of the second phase (the later response seen after a meal). These abnormalities of insulin secretion are reversible after certain forms of bariatric surgery for morbid obesity in patients with type 2 diabetes.

Population studies indicate that the concurrent existence in an individual of both a cause for insulin resistance (usually obesity) and of a relatively low insulin secretory reserve predicts the later onset of type 2 diabetes mellitus.

The differentiation of idiopathic type 2 diabetes and 'secondary' diabetes can be difficult. Secondary diabetes is a term that implies that another disease process has caused the diabetes (see [Box 15.4](#)). While there is good understanding of the natural history and approach to treatment in type 2 diabetes, there is less for secondary diabetes: occasionally, the diabetes can be improved significantly by treating the primary condition.

Genetic factors in type 2 diabetes mellitus

Family studies suggest that type 2 diabetes mellitus is strongly inheritable. Concordance rates for identical twins exceed 90%. Some racial groups have a very high incidence of type 2 diabetes. Notable examples of this include the Pima Indians of Arizona and Pacific Islanders, with prevalence rates of up to 50%. In the UK, the prevalence of type 2 diabetes in people of South Asian extraction is approximately two or three times that in people of European origin. African-Caribbean people show an intermediate prevalence. The natural history of type 2 diabetes and its propensity to give rise to long-term complications varies between races (examples being the relative lack of diabetic foot disease in British Asians and the high prevalence of diabetic nephropathy among those of African-Caribbean descent).

In the majority of patients with type 2 diabetes, the pattern of inheritance suggests a polygenic disorder, with an important role for environmental factors such as obesity and a low level of exercise.

Molecular biological techniques have not yet shown type 2 diabetes to be consistently associated with any abnormalities of the DNA coding of insulin, the insulin receptor or glucose transporter peptides, except in a small percentage (<1%) of cases. Abnormalities of the glucokinase gene and of certain hepatic nuclear factor genes have been shown to cause some cases of maturity onset diabetes of the young (MODY) (see [p. 292](#)), but not typical type 2

diabetes mellitus. More recent genome wide association studies have identified several predisposing genetic factors; most of these appear to relate to β -Cell function, but interestingly the strongest factor yet identified (the *FTO* gene, which is expressed in the hypothalamus and thought to be involved in body weight regulation), appears to increase diabetes risk by predisposing to a higher body weight; homozygotes for the predisposing gene variant are on average 3 kg heavier than those without.

Glucoregulatory defects in type 2 diabetes mellitus

The exact cause(s) of type 2 diabetes mellitus is (are) unknown. Hyperglycaemia is due to elevated hepatic glucose output and, to a lesser extent, failure of skeletal muscle to take up glucose and store it as glycogen. Although many factors are believed to contribute to these two defects, they can broadly be categorized into three groups: β -Cell deficiency/dysfunction, insulin resistance and abnormalities of non-insulin-mediated glucose disposal. It is also perhaps useful to indicate that, although only abnormal concentrations are observed in clinical practice, fluxes can be measured in research studies, and some of these are indicated in [Table 15.6](#).

Pancreatic β -Cell deficiency/dysfunction in type 2 diabetes mellitus. The importance of β -Cell deficiency varies between different groups and different individuals with type 2 diabetes mellitus. Abnormalities of insulin secretion are present in all patients, but the causes(s) of these defects is/are not yet established. Nor is it yet clear whether an abnormality of insulin secretion is usually the primary aetiological defect or whether other glucoregulatory defects are secondary to this pancreatic defect. Several workers have suggested that β -Cell dysfunction is the primary abnormality of type 2 diabetes, but that β -Cell dysfunction may be more subtle than that seen in type 1 diabetes.

Insulin deficiency has been described in type 2 diabetes mellitus, but may sometimes be due to misclassifying a late onset type 1 patient. Some patients with type 2 diabetes can exhibit both insulopaenia and hyperinsulinaemia (relative to normal weight controls) at different times during a single day. Some, especially obese subjects with mild glucose intolerance, may have hyperinsulinaemia throughout the whole 24h. Other patients, usually

TABLE 15.6 Some comparisons of concentrations and fluxes in obesity/insulin resistance syndrome/early type 2 diabetes

Substance or process	Concentrations	Fluxes
Glucose	Increased fasting concentration	Normal whole body disposal rate Normal whole body glucose production and clearance
Non-esterified fatty acids	Variable	Usually normal fasting total body production Impaired responsiveness to meal
Very low density lipoproteins	Increased concentration	Increased production, increased plasma half-life and increased whole body clearance
Oxidation of fuels	Whole body respiratory quotient (RQ) reflects diet	Whole body RQ reflects diet but RQ is lower in obese subjects

those who are less obese, resemble patients with late onset type 1 diabetes mellitus and have absolute insulin deficiency throughout. Fortunately, the accurate distinction between insulin-deficient type 2 diabetes mellitus and late onset type 1 diabetes mellitus rarely has significant implications for the clinical management of individual patients.

Patients with type 2 diabetes mellitus most frequently show hyperinsulinaemia during fasting, combined with reduced β -Cell reserve, relative to healthy subjects. The time course of insulin secretion in type 2 diabetes is abnormal: subjects typically exhibit relative insulin deficiency during both the early phase of insulin secretion after an oral glucose load or meal and the first-phase insulin response to an intravenous glucose load. This loss of early insulin response to glucose is paralleled by defects in the pulsatility of insulin secretion.

Assuming insulin clearance to be normal in such (hyperglycaemic) type 2 diabetes patients, hyperinsulinaemia implies hypersecretion of immunoreactive insulin compared with healthy (euglycaemic) subjects. However, secretion of insulin by β -Cells is dependent on the prevailing blood glucose concentration as well as several other factors. Thus, the level of insulinaemia can be considered a reflection of β -Cell function only if it is considered in relation to the blood glucose concentration. The hyperglycaemic drive to the islets may compensate or even overcompensate for β -Cell dysfunction, producing near-normal or even supranormal insulin concentrations. In the 1990s, it became clear that conventional radioimmunoassays for insulin failed to distinguish mature insulin from proinsulin and partially split proinsulins. Since the precursor molecules have much less biological activity than mature insulin, conventional radioimmunoassays had probably been overestimating the degree of insulinaemia typically found in type 2 diabetes. However, even modern specific assays do confirm at least some degree of hyperinsulinaemia in obese type 2 patients with milder degrees of hyperglycaemia.

With increasing time, even patients who were hyperinsulinaemic at diagnosis usually become relatively insulin deficient. This group of patients, together with those who are insulin deficient from diagnosis, often need exogenous insulin treatment to maintain near-normal glycaemia. They may then be termed 'insulin-treated' or 'insulin-requiring', but it should be recognized that such insulin-treated patients form a heterogeneous group, very different in character from type 1, insulin-dependent patients.

Amylin. Amylin (also known as islet amyloid polypeptide, IAPP) is a 39-amino acid peptide co-secreted with insulin by β -Cells in all subjects with intact insulin secretion, but not those with type 1 diabetes. The amino acid structure has some homology with calcitonin gene-related peptide. Plasma concentrations of amylin are very low ($<10^{-10}$ molar) in both diabetic and non-diabetic subjects. There is no established physiological role for the peptide in the systemic circulation, but it has been suggested that amylin may have a physiological role in the regulation of insulin secretion within pancreatic islets or some effects on bone metabolism. Possible

pathophysiological roles of amylin include the induction of insulin resistance in skeletal muscle, but this only occurs at pharmacological concentrations. Amylin fibrils (with typical amyloid features of secondary protein structure and insolubility) are deposited in islet cells in conditions of excess insulin secretion (such as insulinoma), and in situations where insulin secretion may initially have been increased but has subsequently declined (such as in old age and type 2 diabetes). The possible role of amylin in the islet damage of type 2 diabetes is under intensive investigation.

Insulin resistance in type 2 diabetes mellitus. In 1970, Berson and Yalow defined insulin resistance as 'a state in which greater than normal amounts of insulin are required to elicit a quantitatively normal [glucose] response'. The concept of insulin resistance had been suggested in the 1930s when Himsworth noted that the same amount of exogenous insulin injected into different diabetic subjects had different antihyperglycaemic effects. Those with lesser antihyperglycaemic responses were labelled insulin insensitive (or insulin resistant). When early assays showed that many patients with type 2 diabetes had high concentrations of circulating insulin as detected by radioimmunoassay, the concept of insulin resistance was reinforced. These patients were hyperglycaemic, and hence by definition relatively insulin deficient, yet they actually appeared to have more immunoreactive insulin than other people so that their true insulin requirement was believed to be larger still. Hyperinsulinaemia with eu- or hyperglycaemia is generally taken to indicate insulin resistance, since hyperinsulinaemia produces hypoglycaemia in subjects with normal insulin sensitivity. The type 2 patients who apparently needed supranormal amounts of insulin were seen to be the same group that Himsworth had found to be insulin resistant.

As insulin has several actions, resistance can take several forms. It appears that some subjects show resistance to its hepatic effects while some show resistance to its effects on skeletal muscle (activation of muscle glycogen synthase by insulin is often defective), and some show resistance to liporegulatory effects, the degree of resistance being different for different actions of insulin. There is no consensus as to the cellular mechanisms underlying insulin resistance in most patients with type 2 diabetes, though several have been suggested (see Fig. 15.2). Some workers have highlighted the competition between lipids and glucose as metabolic fuels (the glucose-fatty acid cycle (Randle cycle) hypothesis): it can be shown experimentally that high circulating concentrations of alternative fuels such as triglycerides, NEFA, lactate and ketone bodies compete with glucose for uptake and that, in their presence, glucose clearance is reduced. Others have suggested that insulin resistance is a consequence of 'cellular satiety', seen whenever intracellular sensors such as uridine diphosphate (UDP)-glucosamine detect excess energy supply, and other workers have implicated specific cellular abnormalities such as reduced numbers of insulin receptors, reduced receptor function, dysfunction of second messenger systems and intracellular antagonists of the effects of insulin.

Abnormalities of non-insulin-mediated glucose disposal in type 2 diabetes mellitus. As discussed above, most glucose clearance from the plasma occurs independently of insulin. This process, largely via GLUT1, is defective in type 2 diabetes, and contributes to hyperglycaemia, although the mechanism is uncertain. Changes in tissue blood flow in diabetes, particularly within skeletal muscle, may also reduce clearance of plasma glucose.

Associations of type 2 diabetes mellitus

The metabolic syndrome and obesity. Type 2 diabetes mellitus often occurs in patients with a syndrome of morphological and metabolic abnormalities together with associated conditions that has been termed the metabolic syndrome (also known as 'Reaven syndrome', 'syndrome X' or the 'hyperinsulinaemia syndrome'). The features include insulin resistance, dyslipidaemia, obesity (particular abdominal obesity) and hypertension among others. Unfortunately, there have been multiple definitions of these syndromes (e.g. different factors, different combinations of factors and different cut-offs for their definition) and no consensus has been reached. However, these syndromes all reflect co-segregation of factors in patients with type 2 diabetes mellitus and those with lesser degrees of glucose intolerance, as well as in individuals with normal glucose tolerance. Subjects with metabolic syndrome, with or without diabetes, have an increased predisposition to atherosclerosis, and subjects with glucose intolerance (type 2 diabetes mellitus or IGT) have an increased predisposition to atherosclerosis mediated by these associated risk factors. However, identifying metabolic syndrome itself has no unique implications for clinical management: this should focus on the management of the obesity (hence insulin resistance) and the individual cardiovascular risk factors. Some workers specifically exclude obese subjects from this syndrome, but there are many features in common between slim subjects with the metabolic syndrome and those who are obese.

The natural history of type 2 diabetes usually involves an evolution from normal glucose tolerance through impaired glucose tolerance (usually accompanied by other features of insulin resistance syndromes) to the onset of frank diabetes, followed by a continuing β -Cell failure leading to the need for exogenous insulin (Fig. 15.5). Obesity, low physical fitness or other hereditary factors can all hasten this evolution.

The morphological associations of type 2 diabetes mellitus include shorter stature (by 1–4 cm compared with non-diabetic subjects), with obesity of the android type (also known as 'apple', upper body, central or visceral obesity) marked by a high waist:hip ratio, low capillary density in skeletal muscle and high ratios of slow to fast twitch muscle fibres. Insulin resistance is a prominent feature of obesity, especially of the android type, even in the absence of diabetes, and obesity powerfully identifies individuals within a population at greatest risk of developing type 2 diabetes. Compensatory hyperinsulinaemia is marked in obesity and contributes to the associated dyslipidaemia and probably to the hypertension.

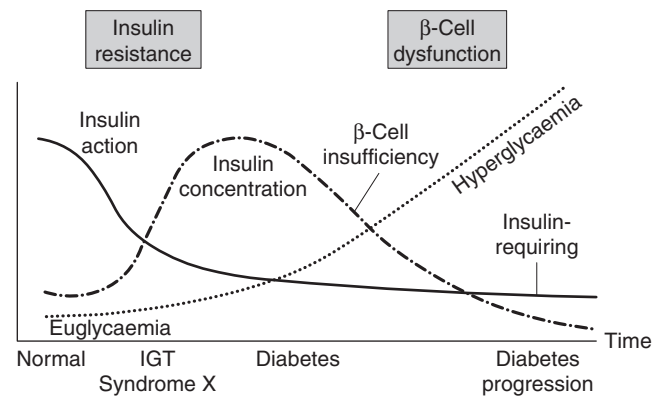


FIGURE 15.5 ■ The natural history of type 2 diabetes. (From DeFronzo et al. *Diabetes Care* 1992; 15:318, with permission).

Hypertension. The association between diabetes and hypertension has been long recognized and is strong. The prevalence of hypertension in obese patients with type 2 diabetes approaches 50% in some series. The typical body habitus of patients with type 2 diabetes is associated with hypertension even in non-diabetic subjects. Although diabetic patients are liable to develop the same secondary forms of hypertension as the non-diabetic population (and renal artery stenosis is commoner in diabetes), most diabetic hypertensive patients have a low renin hypertension that is unlike that in non-diabetic patients with essential hypertension.

Patients with either type 1 or type 2 diabetes and hypertension and characteristically have sodium retention and impaired natriuresis: exchangeable body sodium is increased by an average of 10%. This abnormality is seen even before the development of any clinically detectable complications of diabetes. Possible sodium-retaining mechanisms include hyperinsulinaemia induced overactivity of tubular sodium transporters, increased glomerular filtration of glucose leading to enhanced proximal tubular sodium–glucose cotransport, extravascular shift of fluid with sodium and, in later stages, renal impairment. Plasma renin activity, angiotensin II, aldosterone and catecholamine concentrations are usually normal in glycaemically well-controlled type 1 and type 2 patients. On the other hand, plasma natriuretic peptide concentrations are usually increased, and an exaggerated vascular reactivity to noradrenaline (norepinephrine) and angiotensin II is common even in uncomplicated types 1 and 2 diabetes.

In type 1 diabetes, hypertension is strongly linked with the development of diabetic nephropathy. Although it is uncertain whether this is initially cause or effect, it becomes a vicious circle. There appear to be familial effects, with non-diabetic relatives of diabetic nephropathic hypertensive patients showing defects in ion transport function (erythrocyte Na^+/Li^+ counter transport and leukocyte Na^+/H^+ antiport) and an increased liability to develop essential hypertension.

Data from the UKPDS confirm the clinical impression that hypertension in type 2 diabetes is often refractory to treatment. Typically, a combination of three antihypertensive agents was required to maintain good

control of blood pressure. Furthermore, the same study showed that even a modest reduction in blood pressure reduced the incidence of nephropathy and death. The relationship between blood pressure and the development and progression of diabetic nephropathy was particularly striking.

Dyslipidaemia. Diabetes is strongly associated with abnormalities of lipid metabolism. Several mechanisms are postulated for these associations. The obesity and body fat distribution common in type 2 diabetes is itself linked with dyslipidaemias in non-diabetic subjects. Non-enzymatic glycation of apolipoproteins impairs lipoprotein clearance. Insulin is the principal antilipolytic regulator, through its inhibition of hormone-sensitive lipase. This enzyme breaks down adipose tissue triglyceride and consequently mobilizes fat stores for subsequent utilization. As a result of its regulation by insulin, hormone-sensitive lipase is most active in the fasting state and least so upon feeding. In diabetes, insulin-mediated inhibition is attenuated or lost, so breakdown of fat stores carries on even if food is available. The uncontrolled release of free fatty acids is followed by their uptake by the liver in a simple concentration-dependent manner.

Free fatty acids are metabolized by β -oxidation, but once their concentration exceeds capacity for oxidation they are re-esterified with glycerol to form triacylglycerol (triglyceride) which leads to increased rate of synthesis (and thereby release) of triglyceride-rich very low density lipoprotein (VLDL). Peripheral VLDL-triglyceride clearance may be impaired because insulin is needed to synthesize and secrete lipoprotein lipase, the principal enzyme responsible for clearing VLDL-triglyceride.

Type 2 diabetes mellitus is associated with low plasma concentrations of high density lipoprotein (HDL), which, by acting as an antioxidant, may limit the lipid peroxidation that is one of the factors responsible for atheroma formation. This may contribute to increased cardiovascular risk in diabetes, but high VLDL-cholesterol and elevated triglyceride concentrations may also contribute. Total plasma cholesterol concentration is often normal in type 2 diabetes, but the HDL:LDL and HDL:total cholesterol ratios are usually low. The composition of VLDL changes such that it contains more triglyceride and cholesteryl esters relative to the apoprotein content. The size distribution of LDL changes towards smaller, denser particles. These abnormalities are particularly atherogenic and underline the need to consider the whole lipid profile rather than just total or LDL-cholesterol in the dyslipidaemia of type 2 diabetes. Improved glycaemic control only partly improves these lipid abnormalities.

In type 1 diabetes, poor glycaemic control is typically associated with high plasma concentrations of VLDL-cholesterol, LDL-cholesterol and triglycerides, and sometimes with low HDL-cholesterol. The triglyceride concentration falls in response to improved glycaemic control and the abnormalities of VLDL and LDL usually also improve. In contrast to patients with type 2 diabetes, those with well-controlled type 1 diabetes usually

achieve plasma lipid concentrations similar to those seen in non-diabetic groups; HDL concentrations may even be elevated.

Lipotoxicity and glucotoxicity

Lipotoxicity. Although adipose tissue is now known to have a large number of important functions, related, for example, to thermo-insulation, immunity, fertility and protection of tissues such as the eye in the orbit, its main function is the storage of energy in the form of triglycerides in the postprandial state and the subsequent release of lipid in the form of NEFA in the fasting state. The consequences of accumulation of lipids in lean tissues, as observed in fatless rodents and humans with generalized lipodystrophy, include hepatic steatosis, lipid-induced cardiomyopathy, insulin resistance and type 2 diabetes mellitus. This process is termed lipotoxicity (lipoapoptosis, where programmed cell death occurs) and is reversible in fatless rodents by the transplantation of small amounts of normal adipose tissue, but not by transplantation of adipose tissue from *ob/ob* mice, which lack the ability to secrete leptin. In humans with generalized lipodystrophy, long-term treatment with leptin improves insulin resistance, hyperlipidaemia and hepatic steatosis dramatically.

Non-esterified fatty acids can induce insulin resistance in muscle via at least three putative mechanisms: through the glucose-fatty acid cycle, whereby increased intracellular citrate derived from fat leads to diversion of glucose away from the tricarboxylic acid; through an effect of intermediates such as diacylglycerol that increase the activity of protein kinase C activity, thereby phosphorylating and inactivating the insulin receptor and insulin receptor substrate (IRS)-1, and via activation of the NF κ B transcription factor pathway, which has putative vascular effects that might contribute to the observed increase in vascular damage that precedes hyperglycaemia. In the liver, NEFAs inhibit the suppression of glycogenolysis by insulin. Other mechanisms include modulation of adipokines, such as tumour necrosis factor α (TNF α), which could promote insulin resistance, or adiponectin, which reduces it. The question then arises of why it is that so many individuals with obesity and elevated plasma concentrations of NEFA do not develop diabetes. One explanation for this may lie in the fact that NEFAs are potent stimulators of insulin secretion in healthy individuals, an effect that might mitigate the tendency for NEFAs to induce insulin resistance. However, both in subjects with type 2 diabetes and their normoglycaemic first-degree relatives, NEFAs do not induce a sufficient compensatory rise in insulin secretion to overcome the induced insulin resistance. Thus NEFAs may be able to cause diabetes in those with a genetic predisposition to β -Cell dysfunction but not in those with normal β -Cell reserve. Activators of PPAR γ (see p. 315) may alleviate several of these NEFA-induced abnormalities by reducing plasma NEFA concentrations, increasing adiponectin and redistributing fat from visceral to subcutaneous deposits, thus reducing the direct effects on the liver and elsewhere. Lipotoxicity may also contribute to β -Cell dysfunction – accumulation of lipid droplets has been observed in the islets in rodent

models of obesity and diabetes, and these may both interfere with insulin secretion and promote an inflammatory response leading to β -Cell destruction.

Glucotoxicity. It has long been known that acute elevation of plasma glucose concentrations to very high levels is able to induce a state of insulin resistance characterized by impairment of insulin secretion in response to glucose. Elevation of plasma glucose to concentrations just above the physiological range potentiates insulin secretion in both humans and animals when transient and in the presence of a normal β -Cell mass, but reduces it otherwise. Conversely, strict metabolic control is able to induce improvements in both insulin secretion and sensitivity, although not usually to normality. It is likely that multiple mechanisms contribute to this effect, including changes in the K_m of glucose sensing systems such as glucokinase/hexokinase, which may lead to alteration of the dose-response curve of islet cells to blood glucose concentrations, changes in the ratios of proinsulin to insulin secretion and alteration in the functional activity of the membrane sulfonylurea-sensitive K^+ channel. It is likely that the honeymoon period often observed in new onset type 1 diabetes is at least partly attributable to a reduction in glucotoxicity.

Prevention studies in type 2 diabetes

Type 2 diabetes may be prevented, or at least its onset delayed, by relatively modest degrees of weight loss. In four intervention studies, conducted in the USA, Finland, China and India in patients with impaired glucose tolerance, lifestyle interventions resulting in average weight loss of <7 kg over six months, with some later regain, resulted in up to 58% reduction in the cumulative incidence of diabetes in the intervention groups over periods of up to four years. Other studies have shown that adopting a more physically active lifestyle appears to confer useful protection, independently of body weight. For patients unable to lose weight after appropriate lifestyle interventions, treatment with metformin may bring about a modest reduction in the incidence of type 2 diabetes. Other diabetes treatments, including acarbose, thiazolidinediones and the weight loss drug orlistat also reduce progression to diabetes in those a greatest risk.

The role of bariatric surgery in managing type 2 diabetes

Weight loss is the cornerstone of the management of overweight and obese subjects with type 2 diabetes. Increasingly, patients are being considered for bariatric surgery after other methods of weight control have proved fruitless. Whereas the time course and degree of improvement in diabetes are more or less in line with predictions based on the degree of postoperative weight loss resulting from operations that physically limit food intake, such as laparoscopic gastric banding, the effects of gastric and intestinal bypass procedures appear to be quite different. Not only have improvements in glycaemia been reported well before weight loss becomes

apparent, but they appear also to exceed those expected on the basis of the amount of weight lost. It has been proposed that increased secretion of a number of gut peptides with insulinotropic actions such as glucagon-like peptide-1 (GLP-1) and glucose-dependent insulinotropic peptide (GIP), and decreased secretion of orexigenic peptides such as ghrelin may be responsible for the success of bypass techniques, raising the intriguing possibility that operations of this kind have a partly endocrine mechanism of action. Many studies have purported to show 'cure' of diabetes, often associated with restoration of first-phase insulin response, in up to 80% of subjects for as long as 16 years of follow-up. However, more recent prospective studies suggest that the true rate of remission of type 2 diabetes after bariatric surgery is closer to 40%, with those patients with shorter duration of diabetes, and lower requirements for glucose-lowering drugs most likely to remain free of the need for treatment for many years. Long-term studies comparing surgery with optimal medical treatment are really needed to resolve the question as to whether surgery is a better treatment.

Gestational diabetes mellitus

The study of diabetes occurring in association with pregnancy has been characterized (and hampered) by the local adoption of many different criteria for the diagnosis of gestational diabetes, coupled with different recommendations for screening. Such local diversity is probably a consequence of different levels of provision of care for pregnant women, and the balance between the unnecessary labelling of a pregnancy as 'abnormal' or 'high risk' and the avoidance of preventable risk to mother and fetus.

Gestational diabetes mellitus (GDM) was defined by Lowy as glucose intolerance that presents in pregnancy. However, it must be appreciated that this will group together a number of conditions that are associated with a variety of qualitative and quantitative threats to the pregnancy. The offspring of women who have undiagnosed pre-existing type 2 diabetes discovered during pregnancy will, for example, have been exposed to hyperglycaemia and other metabolic abnormalities in the first trimester of pregnancy during which organogenesis occurs. They will therefore be subject to the increased risk of fetal malformation seen in pre-existing diabetes despite the diabetes being labelled 'gestational'. When the glucose intolerance is induced by the pregnancy itself, its onset is typically in the late second or early third trimester, and the risks of major fetal malformations are therefore similar to those in the non-diabetic population.

The reported incidence of GDM varies enormously between populations, not least because of local differences in screening procedures and definition. The recommendations of the UK National Institute for Health and Care Excellence (NICE) for screening and diagnosis of diabetes during pregnancy are shown in [Box 15.5](#).

Glycaemic control during pregnancy should be strictly optimized in order to minimize fetal mortality and morbidity. In patients with type 1 diabetes who are attempting to conceive, similarly strict glycaemic control

BOX 15.5 Recommendations for screening for and diagnosis of gestational diabetes mellitus in accordance with UK NICE guidelines

- Women with one or more of the following risk factors for gestational diabetes should be offered testing for gestational diabetes mellitus
 - Body mass index $>30\text{ kg/m}^2$
 - Previous macrosomic baby weighing $>4.5\text{ kg}$
 - Previous gestational diabetes
 - Family history of diabetes (first-degree relative with diabetes)
 - Family origin with a high prevalence of diabetes:
 - South Asian (India, Pakistan or Bangladesh)
 - African Caribbean
 - Middle Eastern
- Screening for gestational diabetes using fasting plasma glucose, random blood glucose, glucose challenge test and urinalysis for glucose should not be undertaken
- The 2-h 75 g oral glucose tolerance test (OGTT) should be used to test for gestational diabetes and diagnosis made using the criteria defined by the WHO
 - Fasting plasma venous glucose concentration $\geq 7.0\text{ mmol/L}$
- or*
- 2-h plasma venous glucose concentration $\geq 7.8\text{ mmol/L}$
- Women who have had gestational diabetes in a previous pregnancy should be offered early self-monitoring of blood glucose or an OGTT at 16–18 weeks, and a further OGTT at 28 weeks if the results are normal
- Women with any of the other risk factors for gestational diabetes should be offered an OGTT at 24–28 weeks

(together with folic acid supplements) is believed to reduce the incidence of congenital malformations. One large UK study showed, in an unselected population, that the infants of women with type 1 diabetes have a ten-fold greater risk of a congenital malformation and a five-fold greater risk of being stillborn than those from a representative general population. These effects are reduced but not abolished by pre-pregnancy counselling, the use of folic acid and tight glycaemic control prior to conception and throughout pregnancy. Local targets for glycaemic control vary, but most would agree that, at the least, all overnight fasting and preprandial blood glucose concentrations should be $<7\text{ mmol/L}$ during pregnancy. In practice, many adopt targets for overnight fasting concentrations of $<5.9\text{ mmol/L}$ and 1 h postprandial glucose concentrations of $<7.8\text{ mmol/L}$, in both pre-existing diabetes and gestational diabetes. Although the use of oral hypoglycaemic agents appears to be safe, they are at present not licensed for use in pregnancy, although they are recommended by many guidelines (including NICE in the UK), on the basis of a published evidence base supporting the use of both metformin and glibenclamide in pregnancy. Glibenclamide does not appear to cross the placental barrier and thus does not cause neonatal hypoglycaemia, but may cause maternal hypoglycaemia, so should be used with caution.

Despite the above, insulin treatment remains the gold standard in all forms of gestational diabetes, despite the fact that, at the time of writing, no insulin is licensed for such use. As with diabetes not associated with pregnancy, many regimens have been used with success and there is no single recommended regimen. However, the predominantly postprandial nature of the glycaemic derangements in GDM lend themselves to treatment with short-acting insulins at meal times, supplemented if necessary with a long-acting insulin, and such regimens usually prove satisfactory.

In patients with pre-existing diabetes, particularly where complicated by microvascular disease, retinopathy and neuropathy may advance rapidly during pregnancy, and screening by fundoscopy and for microalbuminuria in each trimester is probably warranted. Angiotensin-converting-enzyme inhibitors are teratogenic and should be discontinued during pregnancy.

The timing and mode of obstetric delivery merits consideration in pregnancies complicated by diabetes. Again, there is a lack of clear evidence, but many would advocate intervention at around 38 weeks in cases where diabetes was present prior to pregnancy, to reduce the risk of late placental failure; this occurs in approximately 1–2% of pregnancies, presenting with sudden and apparently unexpected late intrauterine fetal death (sometimes manifest by a sudden improvement in glycaemic control).

Gestational diabetes or delivery of a large-for-dates or macrosomic baby is strongly predictive of subsequent type 2 diabetes, and many recommend screening such women for type 2 diabetes mellitus six weeks after delivery, and periodically (every 1–2 years) thereafter.

Maturity onset diabetes of the young (MODY)

This syndrome is a rare ($<5\%$ of childhood diabetes and $<1\%$ of all type 2 diabetes) form of type 2 diabetes, inherited in an autosomal dominant fashion and typically presenting at a young age. It is sometimes known as 'Mason-type' diabetes after the index family. Strict diagnostic criteria for the condition include a diagnosis of type 2 diabetes (with preservation of C-peptide response) with onset before the age of 25 years in at least two family members, with evidence of vertical transmission through at least three generations. In clinical practice, it may not always be possible to ensure that all these criteria are met. Thus, many patients with a probable diagnosis of MODY will not fulfil the criteria by virtue of later age of onset (at least 25% present later than 25 years of age), the occurrence of new mutations and incomplete family histories.

All the six known MODY subtypes are characterized by defects in insulin secretion rather than in insulin action, although, in some, a degree of insulin resistance may supervene, probably as a result of glucotoxicity. They have been characterized as being due to mutations in genes encoding glucokinase (MODY 2) or a range of transcription factors that result in defects in insulin secretion. In patients with MODY1, abnormalities in hepatocyte nuclear factor 4a (HNF-4A) result in impaired or absent first-phase insulin

secretion, reduced peaks of insulin secretion, hyperglycaemia, a degree of secondary insulin resistance reversible by good metabolic control and fully susceptibility to the complications of diabetes expected for the degree of exposure to hyperglycaemia. In contrast, patients with MODY 2 have mild fasting hyperglycaemia, as a result of the defect in glucose sensing that correlates with the degree of functional inactivation of the enzyme. MODY 2 is associated with low risk of microvascular complications of diabetes and thus often does not require treatment. The condition often only comes to light as a chance finding. These patients are characteristically highly sensitive to sulfonylurea treatment, if used. As predicted by the actions of glucokinase in the liver, hepatic glycogen production is reduced and gluconeogenesis increased in the postabsorptive state in this condition.

Mutations in the *HNF-1A* gene that encodes hepatocyte nuclear factor 1a (HNF-1A) (MODY 3) account for up to two-thirds of all MODY diagnoses. This condition has a variable age of onset, and progressive β -Cell dysfunction that leads to potentially severe hyperglycaemia and similar microvascular complications to those encountered in type 2 diabetes. Abnormal HNF-1A function in the proximal renal tubules may lead to incomplete reabsorption of glucose, phosphate and amino acids, but nephropathy appears no more common than would be expected for the degree and duration of hyperglycaemia. These patients are characteristically highly sensitive to sulfonylurea treatment, and there are several cases reported where patients have been misdiagnosed as type 1 diabetes, treated with insulin (sometimes for many years) but later transferred safely to sulfonylurea treatment. MODY 4 results from abnormalities of IPF-1, which plays a role in pancreatic development: patients may present with varying degrees of hyperglycaemia at variable ages of onset. MODY 5 is caused by defects in HNF-1B, which plays a role in renal development as well as β -Cell function. It is characterized by renal cyst formation (which may lead to chronic kidney disease) and variable degrees of hyperglycaemia. MODY 6 is associated with abnormalities of NeuroD1, which plays a role in pancreatic development.

Clinicians need to be aware of the possibility of MODY in patients with a strong family history of early onset type 2 diabetes, and consider genetic testing. Other tests are also being developed, for example low plasma concentrations of C-reactive protein (measured with a highly sensitive assay) can be highly predictive of MODY 3 and may eventually be incorporated into diagnostic algorithms.

Secondary diabetes

Several diseases may cause diabetes or glucose intolerance. Most of these are readily diagnosed, perhaps with haemochromatosis and chronic pancreatitis being the least readily apparent. Clinical concern is frequently raised about Cushing syndrome when an obese subject presents with glucose intolerance, and a number of studies have suggested that up to 4% of patients with type 2 diabetes are in fact suffering from Cushing syndrome. Pseudo-Cushing syndrome, either idiopathic or related to alcohol excess, is a more common association.

Tropical diabetes. Fibrocalcific pancreatic diabetes is a form of secondary diabetes that occurs in the tropical developing countries. The condition has typical clinical features, pancreatic calcification being characteristic, but its aetiology has not been established. There was previously considered to be another form of tropical diabetes termed protein-deficient pancreatic, or J-type, diabetes, but this is not included in current classifications.

Alcohol-related and pancreatic causes of diabetes.

Chronic pancreatitis accounts for <1% of all diabetes in the developed world. Alcohol is the usual cause. Most patients give a history of long-term (>5 years) excessive alcohol consumption with recurrent episodes of mild abdominal pain. Hospitalization during these episodes of pancreatitis is unusual. There may or may not be pancreatic calcification, but sclerosis of islets occurs. Insulin secretion is reduced, causing diabetes, but patients are highly sensitive to exogenous insulin therapy (perhaps because glucagon secretion is also reduced). Chronic heavy alcohol consumption may also induce severe exocrine pancreatic failure and, consequently, steatorrhoea. Pancreatic resection, for example for cancer, is a rare cause of diabetes.

Acute pancreatitis is associated with transient glucose intolerance resulting from reduced insulin secretion and the insulin resistance of systemic illness. Acute haemorrhagic pancreatitis occasionally produces such damage to the pancreas that permanent diabetes mellitus results.

Mild (subclinical) abnormalities of exocrine pancreatic function and reductions in pancreatic lipase and immunoreactive trypsin are seen in 20–70% of diabetic patients. The cause for these abnormalities is not established, although the pancreas is smaller in patients with diabetes (especially type 1 diabetes). Changes in insulin secretion, glucagon secretion and autonomic neural function may also play a role in exocrine pancreatic dysfunction.

Haemochromatosis. Iron storage disease, whether owing to familial defects or repeated iron infusion, results in liver damage that may progress to cirrhosis (see Chapter 14) and β -Cell damage, as well as damage to other endocrine and non-endocrine tissues (e.g. cardiac muscle). Diabetes is a common result of such problems. Glucose tolerance improves with treatment of the iron overload, but the more severe the damage, the lower the probability of recovery with treatment. Screening diabetic clinic populations typically reveals 0.1–0.5% of type 2 diabetes mellitus to be associated with undiagnosed iron storage disorders. However, modestly elevated concentrations of ferritin are commonly associated with type 2 diabetes, particularly at the time of diagnosis, and do not of themselves necessarily suggest the diagnosis of haemochromatosis.

Endocrine disorders. In many of the conditions mentioned in this section, insulin resistance occurs but, in most subjects, β -Cell reserve is sufficient to produce compensatory hyperinsulinaemia so that, although there is mild glucose intolerance, full-blown diabetes mellitus

is unusual. The unrecognized development of an endocrine condition in a known diabetic patient may present as worsening glycaemic control or recurrent ketoacidosis. It is more common for endocrine disease to exacerbate or unmask pre-existing glucose intolerance rather than for it to be the sole cause of diabetes.

Polycystic ovary syndrome is a common condition, associated with obesity, insulin resistance and glucose intolerance or frank type 2 diabetes (see below and Chapter 22).

Active thyrotoxicosis is associated with glucose intolerance in 30–50% of patients, but this rarely extends to frank diabetes. The mechanism causing this may vary between patients. Hepatic glucose production is often increased; increased gastrointestinal motility may exaggerate postprandial hyperglycaemia; insulin clearance is increased, and hypersensitivity of β -adrenoreceptors may mediate the elevated NEFA concentrations found in hyperthyroidism.

The increased incidence of impaired glucose tolerance in patients with hypothyroidism is well established, although the mechanism for it is uncertain.

Hyperprolactinaemia causes insulin resistance and glucose intolerance, which is reversed by treating the primary condition; symptomatic diabetes is rare.

The hypercortisolaemia of Cushing syndrome, whatever the cause, results in increased hepatic gluconeogenesis and hepatic glucose output while skeletal muscle becomes insulin resistant. Glucocorticoids increase lipolysis and protein catabolism, increasing the circulating concentrations of alternative fuels and reducing glucose clearance. Glucose intolerance is found in 80–90% and frank diabetes in 15–20% of patients with Cushing syndrome. Treatment with glucocorticoids is a common cause of secondary diabetes.

Conn syndrome was originally described as including glucose intolerance. However, this is usually mild and only a small minority of patients has even moderate glucose intolerance. The cause is thought to be impaired insulin secretion owing to severe potassium depletion.

Growth hormone is a counter-regulatory hormone. Acromegaly frequently causes glucose intolerance (seen in 60–70% of cases) and even symptomatic diabetes (in 6–25%). Successful treatment of acromegaly usually improves glucose tolerance. Isolated growth hormone deficiency is also associated with glucose intolerance, probably as a result of the insulin resistance that results from the decrease in lean body mass and the increase in abdominal fat and a blunted insulin response to glucose. Diabetic microvascular disease is rare in this group.

Phaeochromocytoma is associated with multiple abnormalities of glucoregulation, adrenaline (epinephrine) having a greater hyperglycaemic effect than noradrenaline (norepinephrine). Catecholamines stimulate hepatic glycogenolysis and hepatic glucose output, and inhibit insulin secretion via α -receptors; β -receptor effects include promotion of adipose tissue lipolysis, increased skeletal muscle glycogenolysis and reduced skeletal muscle glucose uptake. Both α - and β -receptors augment glucagon secretion. Improvement of glucose tolerance occurs within a few weeks of successful surgical resection.

Hypercalcaemia and hypophosphataemia, as seen in primary hyperparathyroidism, reduce peripheral insulin sensitivity, probably through effects on intracellular second messenger systems and glucose uptake, and are associated with hyperinsulinaemia, although glucose tolerance is rarely impaired.

Gastrointestinal endocrine tumours of several types are associated with glucose intolerance. Glucose intolerance is a cardinal feature in patients with glucagonomas. Zollinger–Ellison (gastrinoma) syndrome; Verner–Morrison or watery diarrhoea hypokalaemia, achlorhydria (WDHA or VIPoma) syndrome; carcinoid syndrome; polyneuropathy, organomegaly, endocrinopathy, monoclonal gammopathy, skin changes (POEMS) syndrome and somatostatinoma, all have glucose intolerance as a frequent, and often florid, feature.

Autoimmune hypoadrenalism is associated with type 1 diabetes mellitus but not causative: indeed, adrenal failure is a cause of hypoglycaemia.

Iatrogenic diabetes. Treatment with steroids is the most common form of iatrogenic diabetes. Some frequently used drugs, including thiazides and β -blockers, especially in combination, may worsen glucose tolerance, their effects being most clinically relevant in diabetic patients. Other drugs may cause diabetes via effects on lipid metabolism (e.g. protease inhibitors), via weight gain or through mechanisms that are not fully understood (e.g. atypical antipsychotic agents).

Rare conditions associated with glucose intolerance. Several congenital conditions with varying degrees of impairment of glucose intolerance are listed in [Table 15.7](#). In most of these conditions, the diabetes is not the most pressing clinical problem.

Severe insulin resistance. Most of the endocrine causes of glucose intolerance are the result of circulating insulin antagonists causing insulin resistance. A variety of other medical conditions, although rare, may be associated with severe resistance to insulin.

Acanthosis nigricans is characterized by the presence of velvety brown hyperkeratotic lesions on the neck, axillae and groins. There is a well-recognized, but relatively rare association with malignancy. Acanthosis nigricans not associated with malignancy may be classified into two types, both associated with insulin resistance and obesity. Type A is a variant of polycystic ovarian syndrome where the skin changes are marked; additional features include hirsutism, polycystic ovaries, virilization, coarse features and early accelerated growth. The cause of the insulin resistance in this type has not yet been determined. Several factors may be involved: reductions in insulin binding, receptor number and receptor kinase activity, and post receptor defects have each been reported in some patients. Patients with type B acanthosis nigricans tend to be older and are usually female. They have markers of autoimmune disease including hypergammaglobulinaemia, proteinuria, hypocomplementaemic nephritis, leucopenia, arthralgia, alopecia, enlarged salivary glands and positive antinuclear and anti-DNA antibodies. They have reduced insulin binding to monocytes *in vitro*, owing to

TABLE 15.7 Congenital conditions with associated glucose intolerance

Condition	Main features
Klinefelter syndrome	Skeletal, gonadal (infertility)
Turner syndrome	Skeletal, gonadal (infertility)
Down syndrome	Neurological
Friedreich ataxia	Neurological
Ataxia telangiectasia	Neurological
Dystrophia myotonica	Neurological
Huntington chorea	Neurological
DIDMOAD	Diabetes insipidus, diabetes mellitus, optic atrophy and deafness
Refsum disease	Neurological
Alstrom syndrome	Neurological
Laurence-Biedl-Moon syndrome	Neurological, obesity
Prader-Willi syndrome	Neurological, obesity
Cystic fibrosis	Respiratory, pancreatic (malabsorption)
Werner syndrome	Growth retardation, premature senility
Fetal rubella syndrome	Neurological, cardiac
Lipodystrophy (various forms)	Metabolic, renal, vascular, hepatic
Glycogen storage disorders	Hepatic/myopathic

the presence of an autoantibody against the insulin receptor. Ataxia telangiectasia may show some overlap with the features of type B acanthosis nigricans.

Leprechaunism is a rare congenital condition with typical facies, lipodystrophy, cliteromegaly, hirsutism and acanthosis nigricans. It is usually fatal. Affected children are severely resistant to exogenous insulin. They produce large amounts of normal insulin endogenously. Several cellular defects have been described that produce the phenotype of leprechaunism. Most patients have defective kinase activity of the insulin receptor, although some show defective insulin receptor formation and others are unable to recycle insulin receptors back to cell membranes after insulin binding. The common features of acanthosis nigricans and virilism are probably an effect of stimulation of IGF receptors by high concentrations of insulin.

Lipodystrophy occurs in local and generalized forms; it is usually familial. There are several associated features but these vary from family to family. Acanthosis nigricans, hepatosplenomegaly, nephritis and hyperlipoproteinaemia have all been reported. Likewise, the cellular defect varies, with some families having reduced insulin binding and some having reduced receptor numbers. Lipodystrophy is also frequently seen as a complication of treatment of acquired immune deficiency syndrome (AIDS), particularly with protease inhibitors. This may be severe enough not only to be disfiguring, but also to cause severe hyperlipidaemia and diabetes mellitus. There is something of a paradox in the fact that type 2 diabetes is associated with increased visceral fat depots but also with lipodystrophy. This is probably explained by the role that adipose tissue has in storing fatty acids as triglycerides. In conditions where this process is impaired,

circulating concentrations of NEFA are increased and may affect hepatic insulin sensitivity in much the same way that they do when there is excess visceral adipose tissue accumulation in obesity.

Anti-insulin antibodies. A previously common form of insulin resistance was iatrogenic, owing to the induction of anti-insulin antibodies (also known as insulin antibodies) by the use of exogenous pork or beef insulins for the treatment of diabetes. Treatment with human insulin *ab initio* results in very few (if any) patients developing such antibodies. Anti-insulin antibodies bind exogenous insulin, reducing its plasma concentration and action. Clinically, the presence of significant anti-insulin antibodies manifests as either a high insulin requirement for the achievement of glycaemic control or with altered insulin kinetics (usually delayed insulin action). The antibodies may be detected by a variety of assays. Treatment is usually by changing to insulin analogues, human or (very rarely) sulfated animal insulins.

Insulin-binding autoantibodies are also seen as a result of autoimmune disease or haematological malignancy and may cause profound insulin resistance, although ketosis rarely occurs as the binding of insulin to IGF receptors is not affected and appears to allow sufficient suppression of lipolysis to prevent excess ketogenesis. These autoantibodies may be an indication for steroid therapy, although high doses of exogenous insulin may be satisfactory.

Cirrhosis. Glucose intolerance occurs in most, and fasting hyperglycaemia in 30% of patients with cirrhosis, unless the latter is due to idiopathic haemochromatosis, when diabetes is common. There are multiple gluco-regulatory defects in cirrhosis including insulin resistance, hyperinsulinaemia, impaired insulin clearance, elevated growth hormone concentrations and hyperglucagonaemia. However, in clinical practice, hypoglycaemia resulting from reduced storage, synthesis and mobilization of glucose typically dominates the clinical picture in a patient with advanced hepatic cirrhosis.

ENDOCRINE ASSOCIATIONS WITH DIABETES

Several endocrine conditions occur with increased frequency among diabetic populations. Endocrine conditions that cause secondary diabetes have been discussed above.

Abnormalities of glucagon secretion appear to be a consequence of diabetes. In normal subjects, insulin inhibits glucagon release and glucagon concentrations fall during hyperglycaemia. Hyperglucagonaemia is usual in both types 1 and 2 diabetes, and glucagon concentrations in diabetic patients are relatively insensitive to changes in glucose concentrations. The mechanisms for this are unknown but may relate to de-repression of glucagon release by the insulin deficiency of diabetes. In longstanding type 1 diabetes, the glucagon response to hypoglycaemia may be lost, which may contribute to poor recovery from hypoglycaemia.

Some other endocrine disorders show an association with diabetes, but the link is shared susceptibility rather

than causal; these include autoimmune hypoadrenalism, Graves and Hashimoto diseases. The development of thyrotoxicosis owing to Graves disease in a patient with diabetes may underlie subsequent loss of diabetic control and weight loss. The prevalence of primary hypothyroidism is reported to be 4–17% in type 1 diabetes.

Type 1 diabetes is statistically associated with coeliac disease (up to 6% of patients may be affected), and both are related to HLA-DR4. Guidelines recommend screening for coeliac disease in patients with type 1 diabetes. The combined dietetic requirements of the two conditions may be onerous for patients. The diarrhoea of unrecognized coeliac disease may be mistaken for diabetic diarrhoea, and coeliac disease should be considered as a cause of newly developing hypoglycaemia in a patient whose glycaemic control was previously poor.

DIABETES, NUTRITION AND GROWTH

In well-nourished children with well-controlled type 1 diabetes, skeletal growth is normal. Poor glycaemic control is associated with elevated growth hormone concentrations but reduced IGF concentrations and is thus a growth hormone-resistant state. It is thought that the latter is related to the reduced growth velocity seen in uncontrolled diabetes.

As a group, patients who develop diabetes as children tend to be more centrally obese than their non-diabetic siblings. It is not yet clear whether this problem is lessened by better glycaemic control.

It is important to appreciate that children with diabetes need a diet that permits healthy growth as well as maintaining good glycaemic control.

MECHANISMS OF DIABETIC TISSUE DAMAGE

Introduction

Both types 1 and 2 diabetes are associated with various long-term complications, also known as diabetic tissue damage. The acute metabolic complications of diabetes are discussed in Chapter 16. Diabetic tissue damage includes a range of 'microvascular complications' (e.g. retinopathy and nephropathy), 'macrovascular complications' (ischaemic heart disease, peripheral vascular disease, stroke and renal artery stenosis) and some complications that do not fit entirely into either category (e.g. glove and stocking peripheral neuropathy, mononeuritis multiplex, cranial neuropathies, entrapment neuropathies, proximal motor neuropathies, autonomic neuropathy, various sorts of cataract and diabetic cheiroarthropathy). Diabetic subjects have a reduced life expectancy and macrovascular disease is, in most cases, the major determinant of this.

Few organs or tissues are exempt from potential damage in diabetes; approximately half of those diagnosed with type 2 diabetes have evidence of tissue damage at diagnosis. However, only relatively few cell types are known to be vulnerable to direct damage from chronic hyperglycaemia, for example the mesangial cells of the

kidney, vascular endothelial cells, pancreatic β -Cells, Schwann cells and neurons. Many other cells appear to be able to reduce glucose uptake under conditions of extracellular hyperglycaemia and, perhaps for this reason, are relatively resistant to direct hyperglycaemic damage.

The two main clinical subtypes of diabetic tissue damage, macrovascular and microvascular disease, are differentially affected by the cumulative effect of hyperglycaemia. It has been proposed that microvascular tissue damage is the result of hyperglycaemia per se, even if accelerated by other factors. Macrovascular complications are frequently found in epidemiological studies to be associated with insulin-resistant states and hyperinsulinaemia, and their relationship to hyperglycaemia is less clearly established. Data derived from the UK Prospective Diabetes Study (UKPDS) suggest that the risk of microvascular disease increases approximately ten-fold with an increase in HbA_{1c} from 37 to 80 mmol/mol (5.5–9.5%), whereas the risk of macrovascular disease becomes apparent at values of HbA_{1c} below that typically associated with diabetes but increases only by a factor of about two-fold with the same increase in HbA_{1c}.

Pathogenesis

The pathogenetic mechanisms involved are ill understood and, for even the major forms of diabetic tissue damage, it is not easy to determine the role of purely diabetic abnormalities (e.g. hyperglycaemia per se and hypo- or hyperinsulinaemia) as opposed to abnormalities related to diabetes (e.g. hypertension, dyslipidaemia or obesity).

The first cellular mechanism to underlie purely hyperglycaemic tissue damage to be described was increased activation of the polyol pathway. The enzyme aldose reductase, usually involved in the reduction of toxic aldehydes to their respective alcohols, is diverted under conditions of intracellular hyperglycaemia to reducing excess glucose to sorbitol, which is later oxidized to fructose. This process consumes the cofactor NADPH, which is then not available for the regeneration of reduced glutathione, thus rendering cells vulnerable to the effects of oxidative stress. Sorbitol accumulation in the lens may cause cataracts, and has been demonstrated in biopsies from diabetic subjects to occur in the eye, peripheral nerves and renal glomeruli. Unfortunately, trials of aldose reductase inhibitors have conferred only marginal reduction in the progression of neuropathy in people with diabetes. This is either because the pathway is not important, is one of several mechanisms of diabetic tissue damage or is a downstream process whose correction would not be expected fundamentally to alter cellular responses to hyperglycaemia. As a result, these agents are not presently used in routine clinical practice.

The next mechanism to be proposed was mediated by the slow accumulation over time of advanced glycation end products (AGEs), which are irreversibly formed by the non-enzymatic glycation of matrix, cellular and plasma proteins. Glucose and amino acids combine to form unstable Schiff base adducts, which undergo chemical rearrangement over time to form Amadori products and, eventually, AGEs that are chemically stable, irreversibly attached to proteins and able to trap other protein elements by covalent binding

and promotion of cross linking. Advanced glycation end products are thought to cause tissue damage by alterations in the structure and function of the extracellular matrix with accumulation of PAS-positive material (e.g. in the basement membrane of blood vessels), by activation of inflammatory cytokines and growth factors via stimulation of AGE receptors (in part by the NF κ B pathway) and by alteration of cellular genetic material. These changes may contribute to the endothelial dysfunction, basement membrane thickening and increased vascular permeability observed in the diabetic microvasculature.

Another mechanism postulated to be involved in the evolution of diabetic tissue damage is that intracellular hyperglycaemia, by causing increased diacylglycerol (DAG) concentration, may cause further activation of the NF κ B pathway via activation of protein kinase C (PKC). The end result is vasoconstriction and hypercoagulability via increased endothelin-1, TGF- β and plasminogen activator inhibitor 1 (PAI-1) generation and reduced endothelial nitric oxide synthase production.

The fourth major mechanism proposed to result in diabetic tissue damage results from shunting of excessive intracellular glucose via fructose 6-phosphate to glucosamine 6-phosphate, a reaction catalysed by the enzyme glutamine:fructose 6-phosphate amidotransferase (see Fig. 15.2). Glucosamine 6-phosphate undergoes further conversion to UDP N-acetyl glucosamine, which, by binding to serine and threonine residues on transcription factors, leads to increased proinflammatory cytokine activity (TGF- β , PAI-1 and others).

A further mechanism relating to the generation of reactive oxygen species (ROS) and oxidative stress has recently been proposed. Excessive intracellular glucose or NEFAs are thought to stimulate tricarboxylic acid cycle reactions by substrate accumulation. Citrate, formed from NEFA or glucose-derived acetyl-CoA and oxaloacetate, is converted to isocitrate by the enzyme isocitrate dehydrogenase, generating mitochondrial NADH in the process. Excessive generation of NADH leads to ROS formation via increased electron transport along the inner mitochondrial membrane and, perhaps when other means of electron dissipation such as uncoupling proteins are saturated, their donation to oxygen species, generating toxic ROS. β -Cells are thought to be especially vulnerable to the resulting oxidative stress by virtue of a low antioxidant capacity. In hyperglycaemic clamp studies, glutathione infusion has been shown to restore islet cell function (particularly first-phase insulin response), and hyperglycaemia or excessive NEFA-induced endothelial dysfunction may similarly be ameliorated by antioxidant infusion. However, data from various studies do not demonstrate convincing benefits from supplemental vitamin E (an antioxidant) in diabetic or other subjects at high risk of cardiovascular disease.

The precise interrelationship between these proposed mechanisms of diabetic tissue damage, for example whether they may be downstream effects of a common maladaptive cellular response to hyperglycaemia or whether they represent independent pathways, is not yet clear. There is compelling evidence, for example that PKC activation may be stimulated by superoxide ions and AGE formation, thus linking the oxidative

stress, advanced glycation end product and PKC pathways. Ruboxistaurin mesylate (LY333531), an inhibitor of PKC, has shown promise in human and animal trials in slowing progression of, and in some cases reversing, a number of diabetic microvascular complications including nephropathy, retinopathy and neuropathy.

Other aspects of diabetic tissue damage

Although the mechanisms discussed above relate to the specific effects of hyperglycaemia, other factors such as infection, genetic susceptibility, hypertension, vascular disease, hyperlipidaemia and other associated conditions such as hyperuricaemia and gout are important and common cofactors for tissue damage in the clinical setting, and these are discussed briefly below.

Tissue damage is encountered in diabetes as a result of increased susceptibility to infection (e.g. renal scarring from repeated episodes of urinary sepsis). Phagocyte (principally neutrophil) dysfunction appears to be the main predisposing factor; this occurs via a number of mechanisms including abnormal adherence, chemotaxis, phagocytosis, respiratory burst and bacterial killing. It is not known precisely what duration of exposure to hyperglycaemia is required to produce these abnormalities in human subjects: the concentration threshold is probably ~ 11 mmol/L.

High concordance rates between twins for complications of diabetes suggest that inherited factors may influence their development. There is evidence that autoimmune processes contribute to the development of some diabetic complications (e.g. the autonomic neuropathy/iritis syndrome).

There is good evidence that growth hormone deficiency, either coincidental or induced by hypophysectomy, prevents or reverses the onset of microvascular complications. However, the results of trials using pegvisomant, a growth hormone receptor antagonist, have been disappointing.

The role of recurrent hypoglycaemia in diabetic tissue damage, especially neurological damage to CNS and autonomic neurons, is currently being investigated but hypoglycaemia is certainly not a necessary condition for its development.

Although microvascular disease and some other complications may be more characteristic of diabetes, macrovascular disease accounts for most of the excess mortality associated with the condition. Diabetic macrovascular disease is related to, but not identical to, atherosclerosis occurring in non-diabetic patients. In diabetic macrovascular disease, the age of onset is earlier and the prevalence of vascular disease in men and women is approximately equal. Differences in the ranking of the strength of associations of risk markers in diabetic and non-diabetic populations suggest some differences in aetiology: smoking is the main risk factor in non-diabetic populations, but hypertension seems relatively more important in diabetes. Furthermore, diabetic macrovascular disease has some singular clinical features, verging on the pathognomic. For example, there is a strong association with calcification of the vascular media and a predilection for involvement of multiple distal arteries (a pattern that

often precludes surgical intervention) rather than fewer proximal arteries. The well-documented higher mortality rate from myocardial infarction in patients with diabetes (compared with similar patients without diabetes) may be related to this widespread involvement of middle-sized and small arteries. Acute ('stress') hyperglycaemia may also contribute to increased mortality, but trials of strict glycaemic control in the immediate post-infarct period have produced mixed results, possibly because the magnitude of the hyperglycaemic response may also be related to the severity of the infarct.

The causation of the increased macrovascular disease risk observed in diabetes is complex; most of the conventional risk factors for vascular disease in non-diabetic individuals (e.g. smoking, hypertension, age) are associated with accelerated atherosclerosis in diabetic populations. However, there is evidence that hypertension in diabetes is in some ways distinct from essential hypertension (e.g. the relationship between plasma renin concentration and plasma renin activity is altered and renal and endothelial dysfunction may be more pronounced). The lipid abnormality most commonly linked with mortality in diabetes may not be high LDL-cholesterol, as found in non-diabetic subjects, but rather the combination of low HDL-cholesterol, small, dense LDL and hypertriglyceridaemia, as also seen in the metabolic syndrome.

Studies of blood rheology, the glycation and peroxidation of lipoproteins and endogenous thrombotic and thrombolytic mechanisms show widespread abnormalities in diabetes, which may all be relevant to the development of macrovascular disease.

The risk of macrovascular disease (whether of the 'normal' or the 'diabetic' pattern is unclear) is increased in subjects with IGT as well as in those with frank diabetes. Most physicians consider that cardiovascular risk factors in patients with IGT should be managed as if they had diabetes.

CONDITIONS ASSOCIATED WITH INADEQUATELY CONTROLLED DIABETES MELLITUS

Various conditions are associated with poorly controlled diabetes, including gout, osteopenia and hepatic steatosis. Acute diabetic emergencies are discussed in Chapter 16.

The association of gout with diabetes has been recognized for >200 years. The reason for the link is not yet clear, but alterations in renal urate clearance may contribute; high uric acid concentrations are associated with other features of the metabolic syndrome and with increased cardiovascular risk. In some patients, hyperuricaemia may be exacerbated by the use of drugs such as thiazides and β -blockers for the treatment of diabetic hypertension.

Generalized osteopenia is also clearly associated with diabetes mellitus. No specific defect in calcium or phosphate metabolism has been identified and the basis of the association is unclear. Of particular importance is local osteolysis, with loss of bone mineral, which in the feet, can be difficult to distinguish from osteomyelitis, especially

in the context of suspect Charcot foot (see p. 324). The inactivity occasioned by diabetic foot ulcers may accelerate the osteopenia. The thiazolidinedione drugs have been associated with increased fracture risk in diabetes.

Fatty change in the liver is common in obesity and in some cases may be associated with an increased risk of developing diabetes. It is also found in poorly controlled diabetes; it frequently causes elevation of plasma liver enzyme activity, notably of alanine aminotransferase and, less frequently, of other enzymes. Liver ultrasound scans will confirm fatty liver in cases of uncertainty. Historically, this has often been considered a benign process, but more recent data suggest that many cases of idiopathic cirrhosis may, in fact, have arisen in fatty livers. There appears to be a spectrum of non-alcoholic fatty liver disease from simple steatosis, which represents the benign end of the scale with little propensity to progress, through various grades of inflammation and fibrosis (steatohepatitis), to frank cirrhosis. There are few good data on selection of patients for biopsy and on measures to prevent progression from simple steatosis to more serious degrees of steatohepatitis. However, most authorities advocate measures to effect good glycaemic control, to treat dyslipidaemia (fibrates may have a particular role) and to improve insulin resistance by exercise and weight loss. Several recent studies have demonstrated beneficial effects from the thiazolidinedione group of drugs, probably as a result both of insulin sensitization and also their effects on NEFA metabolism and fat distribution. This topic is discussed in detail in Chapter 14.

Low concentrations of circulating magnesium have been reported during treatment of ketoacidosis. Hypomagnesaemia may be more common in outpatients with diabetes mellitus than in non-diabetic subjects. However, this suggestion has not been confirmed and there is no evidence that plasma magnesium should be checked routinely in patients with diabetes.

Hyponatraemia may arise in diabetes through several mechanisms. Chlorpropamide and tolbutamide (and perhaps other sulphonylureas) may cause water retention through potentiation of the action of antidiuretic hormone (see Chapter 4). Artefactual hyponatraemia may arise from the hyperlipidaemia (hypertriglyceridaemia) of ketoacidosis or poor glycaemic control. Compensatory hyponatraemia is often seen in poorly controlled diabetes, perhaps as an osmoregulatory response to high glucose concentrations. This effect may be particularly significant in the management of acute diabetic emergencies and is, in effect, a translocational hyponatraemia brought about by the movement of water from the intracellular to the extracellular space as a consequence of the increased plasma osmolality in hyperglycaemic conditions. Plasma sodium concentration should fall by ~ 1.6 mmol/L for every 5 mmol/L that plasma glucose exceeds 5 mmol/L. If the measured sodium differs significantly from this 'corrected' value, a cause should be sought, and if it lies well outside the reference range, fluid management may have to be adjusted accordingly. To what extent, if at all, osmotic tissue damage contributes to the long-term complications of diabetes is at present unclear.

Diabetic kidney damage most commonly presents as hypertension or (micro)-albuminuria in patients under

regular review. Late presentations include chronic kidney disease, nephritic syndrome, nephrogenic diabetes insipidus (sometimes associated with papillary necrosis), renal tubular acidosis (type 4) or toxicity from renally-excreted drugs.

BIOCHEMICAL MEASUREMENTS IN DIABETES MELLITUS

Glucose measurements

Measurements of blood glucose concentrations are fundamental to the diagnosis and management of diabetes mellitus. Plasma contains more glucose (per unit volume) than erythrocytes, so criteria for diagnosis of diabetes reflect these differences between samples (see Table 15.4). Since glucose enters the circulation via the hepatic vein and is cleared peripherally, arterial concentrations are higher than in venous blood samples, with capillary blood having intermediate values.

Over the last decade, home and bedside blood glucose monitoring by patients or their attendants has become widespread, using dry reagent enzymatic reactions or electrochemical methodologies on test strips. Although such techniques represent the only current practical approach to long-term self or nurse monitoring, they are prone to multiple errors if misused (especially by inadequately educated patients or healthcare professionals) and should not be relied upon for clinical decisions in seriously ill patients. They require appropriate calibration and proper technique (e.g. prior washing and drying of hands, correct sample volume and correct test strip for the metre used). Recent advances in testing include electronic transfer of results to databases including the electronic patient record, and alternative site (e.g. forearm) sampling to avoid the pain and repeated trauma associated with fingertip sampling (although forearm glucose concentrations may differ from fingertip concentrations when blood glucose concentrations are rapidly changing, e.g. during exercise or in the postprandial state). Other developments include collection of interstitial fluid by reverse iontophoresis for automated glucose analysis every 10 min, and continuous glucose monitoring systems in which a disposable subcutaneous glucose sensor that can function for several days is monitored via a wired or wireless system. This permits collection of a detailed profile throughout a 24 h period that may yield information that is hard to gather using conventional testing. It is particularly useful overnight, e.g. in patients with suspected nocturnal hypoglycaemia. Glucose concentrations determined by this method lag behind blood glucose by about 10 min, and the method is limited by the development of local skin irritation. The systems are expensive and usually only used for short-term monitoring, although they are sometimes advocated for longer periods of use in patients attempting to achieve tight control of glucose concentrations. The ultimate development of this would be a closed loop system that monitors glucose and calculates the correct amount of insulin to infuse via a pump, but this has not yet gone beyond short experimental studies.

Table 15.8 indicates some factors that may interfere with blood glucose analysis by laboratory or point of care testing methods.

Measurements of urinary glucose concentration are cheap but prone to often correlate poorly with the degree of glycaemic control. Problems include variation in the tubular threshold for glucose (e.g. increased in diabetes, reduced in pregnancy, subject to interindividual variation and altered by a range of drugs and renal tubular disorders) and the fact that the plasma glucose concentration that corresponds to the renal threshold (around 10 mmol/L) is higher than the level that clinical studies (UKPDS and Diabetes Control and Complications Trial, DCCT) have demonstrated to be the target range for blood glucose concentrations. Furthermore, urine glucose concentration is affected by fluid intake and urine concentration and provides information about the entire period since last voiding rather than giving an indication of the current level of glycaemia (unless a second voided specimen is used). The dry reagent stick chemistries used are imprecise at lower levels of glycosuria and may be affected by some drugs. Their use is declining in the developed world; however, they may be more acceptable than blood testing to some patients. They provide an adequate index of glycaemic control in some elderly patients with type 2 diabetes, controlled on diet alone or on small doses of oral hypoglycaemics, although HbA_{1c} (see below) is now the preferred method of monitoring for most of these patients.

Testing for ketones

Urine testing for ketones has an important place in the management of diabetes, particularly for patients with type 1 diabetes, who should generally be instructed in its uses and limitations. Ketone testing is particularly important when metabolic control is threatened by intercurrent illness or stress or when blood glucose concentrations are persistently high (e.g. >15 mmol/L), particularly when accompanied by symptoms compatible with ketoacidosis such as nausea, vomiting or abdominal pain. It must be appreciated that the normally approximately equimolar ratio of the major ketones, β -hydroxybutyrate and acetoacetate, may increase to approximately 6:1 during diabetic ketoacidosis (DKA), reducing during recovery. Furthermore, in recovery from DKA, excretion of urine ketones may continue long after blood acid-base balance is restored to normal. Thus, since all commercially available test sticks use nitroprusside-based reactions, which measure only acetoacetate, tests may become more positive as the patient recovers and remain positive for hours or even days after resolution of the illness. It should also be appreciated that positive results may be found in a significant proportion of normoglycaemic individuals (particularly pregnant women) when testing first-voided urine specimens after an overnight fast. Certain drugs, such as captopril and levodopa, may cause spuriously positive results, while ascorbic acid intake may cause false negative ones.

Enzymatic methods detecting β -hydroxybutyrate in blood using a stick test that can be performed by the patient or at the bedside have a number of theoretical

TABLE 15.8 Factors that may interfere with the biochemical analyses used in diabetes

Substance assayed	Biological fluid	Assay method	Interference from	Error induced
Glucose	Blood	Any	Delay in sample separation, especially if not cooled and in fluoride	Erythrocytes consume glucose giving falsely low value
Glucose	Blood	Glucose oxidase (e.g. YSI (Yellow Springs®))	Paracetamol	Reads high
Glucose	Plasma	Spectrophotometry	Haemolysis or hyperlipidaemia	Affects linearity of response
Glucose	Urine	Any	Bacteriuria	Consumes glucose
Glucose	Urine	Glucose oxidase (e.g. Multistix®, Combistix®)	Detergent (e.g. from cleaning container)	False positive
Lactate	Blood	Any	Reducing substances (e.g. vitamin C), ketones	False negative
Acetoacetate	Blood or urine	Rothera based (e.g. Ketostix®)	Delay in sample separation, especially if not in fluoride	Lactate produced in collection tube
Fructosamine	Blood	Spectrophotometry	Sweat from touching pipettes etc.	Sweat may contain high concentrations of lactate
HbA _{1c}	Blood	Any	Phenolphthalein	False positive
HbA _{1c}	Blood	Electrophoresis, ion-exchange chromatography	Hypertriglyceridaemia, bilirubin	Affects linearity of response
Albumin	Urine	Dye binding	Shortened red cell life span	Lowers value
			Haemoglobin genetic variants	Unable to distinguish glycosylated haemoglobin from haemoglobin variants
			Many detergents, salicylates	False positive

advantages over urine testing. β -Hydroxybutyrate is the most relevant ketone in ketoacidosis, and the use of a metre with the stick permits its quantitative measurement, thus enabling more accurate diagnosis and monitoring of this disorder. The sticks usually also have a longer shelf-life than the urine tests, which is an advantage given that most patients will need to test for ketones only occasionally. Blood ketone measurement is now recommended in some guidelines (e.g. Joint British Diabetes Societies guideline for the management of diabetic ketoacidosis), either by stick testing or using rapid automated laboratory methods, although it is not yet in widespread use.

Oral glucose tolerance test

The oral glucose tolerance test (OGTT) is the reference method for the assessment of glucose tolerance, despite the notoriously poor reproducibility of the test ($CV = 50\%$ for 2 h blood glucose). Most diabetic patients are diagnosed on the basis of symptoms, examination and random or fasting plasma glucose concentrations without recourse to an OGTT. However, WHO and UK guidance is that patients with intermediate fasting glucose concentrations (impaired fasting glycaemia) should undergo formal glucose tolerance testing. The test also has a particular place in the diagnosis of gestational diabetes mellitus, where fasting plasma glucose concentrations lack diagnostic sensitivity. Historically, there has been divergence of opinion as to the dose of glucose to be used. Current WHO recommendations are that this should be 75 g of anhydrous glucose (not glucose monohydrate

which is 10% water by weight). For three days before the test, the subject should be on an unrestricted weight-maintaining diet, with at least 150 g carbohydrate per day, and should exercise normally. The subject should fast overnight for at least 10 h, and should remain seated and not smoke during the test. Oral glucose tolerance tests are not recommended for subjects with fasting plasma glucose ≥ 7.1 mmol/L or hospitalized, acutely ill or immobile patients. Interpretation may be difficult in subjects taking β -blockers, diuretics, nicotinic acid or high doses of glucocorticoids.

Tests of recent glycaemic control

Several tests are available that reflect prior glycaemic control: pre-eminent among these is glycosylated haemoglobin. There are multiple potential glycation sites but the principal one is the β -chain terminal valine residue. Glycation at this site, followed by an Amadori rearrangement to a stable adduct, forms HbA_{1c}. Since the usual lifespan of an erythrocyte is 120 days and the erythrocyte membrane is freely permeable to glucose, glycosylated haemoglobin concentration reflects the glycaemic control over the preceding 120 days, although with substantial weighting towards a shorter time (see below). Many different methodologies, including cation-exchange chromatography, electrophoresis, isoelectric focusing, affinity chromatography and immunoassay can be used to measure glycosylated haemoglobin after its separation from non-glycosylated haemoglobin. All HbA_{1c} assays should be traceable to the reference

method developed by the International Federation for Clinical Chemistry and Laboratory Medicine (IFCC).

The interpretation of glycated haemoglobin measurements will be affected by any coincidental condition that reduces the lifespan of erythrocytes, especially haemolytic anaemias. Conversely, iron deficiency anaemia may spuriously increase HbA_{1c} concentrations by increasing erythrocyte lifespan. Other conditions such as raised plasma concentrations of triglycerides and bilirubin, uraemia, and the presence of haemoglobinopathies may cause interference in some assays.

As an approximation, an HbA_{1c} of 50 mmol/mol (6.7%) corresponds with a mean plasma glucose concentration of approximately 9 mmol/L, and each 10 mmol/mol (1%) increase with a 2 mmol/L increase in mean plasma glucose concentrations. Although HbA_{1c} is a measure of long-term glycaemia, it must be appreciated that more recent effects are relatively more weighted so that the 30 days prior to sampling contribute 50% of the sample result, whereas events from 90–120 days prior to the sample being taken contribute a mere 10%. In conditions where glycaemic control and insulin requirements may change from week to week (such as pregnancy) or where the patient has a structurally abnormal haemoglobin, it may be preferable to use an alternative index of glycaemic control.

Fructosamine is the name given to the ketoamine products of protein glycation formed when glucose bound to a variety of proteins by aldimine linkage undergoes an Amadori rearrangement. The major component of fructosamine in plasma is glycated albumin, but other proteins and possibly non-protein components as well contribute to the measured values. Standardization has been difficult, so reference ranges may differ between different laboratories. Fructosamine is relatively simple to measure (using a nitroblue tetrazolium assay); its concentration reflects control over the preceding 10–15 days, but is subject to spurious results in the presence of factors that affect albumin turnover (including diabetic nephropathy). There is an inverse relationship between weight and fructosamine so that fructosamine concentrations are lower than expected for the degree of control in obese patients with diabetes. Whether or not to correct fructosamine for serum albumin concentration (given that the absolute concentration, not the glycated proportion of the protein, is measured) remains uncertain. The clinical utility of fructosamine is further limited by the lack of evidence-based targets, but it has a limited role in patients with disorders such as haemolytic anaemia or haemoglobinopathies in whom HbA_{1c} cannot be used.

Screening for diabetes

It is estimated that 50% of diabetic subjects in the developed world are undiagnosed. Screening has been recommended by NICE in the UK for asymptomatic subjects who are at high risk of developing diabetes (see Fig. 15.6). This involves a two-stage check – the first is a risk assessment using a validated computer-based risk assessment tool (these usually use factors such as sex, age, BMI, waist circumference, family history, history of hypertension and ethnicity to estimate risk, see, e.g. [\[www.diabetes.org.uk/Riskscore/\]\(http://www.diabetes.org.uk/Riskscore/\)\) in those who meet the following criteria:](http://</p>
</div>
<div data-bbox=)

- adults aged 40 and above, except pregnant women
- adults aged 25–39 of South Asian, Chinese, African-Caribbean, black African and other high-risk black and minority ethnic groups, except pregnant women
- adults with conditions that increase the risk of type 2 diabetes (cardiovascular disease, hypertension, obesity, stroke, polycystic ovary syndrome, a history of gestational diabetes and mental health problems. In addition, people with learning disabilities and those attending emergency medical admissions units, vascular and renal surgery units and ophthalmology departments may be at high risk).

High-risk individuals should be offered a blood test for either HbA_{1c} or fasting plasma glucose. For individuals with results of these tests below the diagnostic threshold for diabetes, an HbA_{1c} of 42–47 mmol/mol (6–6.5%), or a fasting glucose of 5.5–6.9 mmol/L indicates high risk. These individuals should be offered advice on intensive lifestyle intervention and re-testing at least once a year.

Similar recommendations have been made by the American Diabetes Association and other organizations.

Tests for insulin resistance

In subjects who require large amounts of insulin to maintain euglycaemia, e.g. >150 units or >1.5 units/kg body weight/24h, insulin resistance is likely to be present (although it is important to ensure that the patient is compliant with therapy). Many such subjects will be morbidly obese and this is usually an adequate explanation. Measurements of fasting insulin, C-peptide and glucose will show inappropriately high concentrations. Haemolysis reduces insulin concentrations, so care should be taken to avoid this, and blood samples should be kept cool (4°C) until the plasma can be separated. Normal fasting insulin concentrations are up to 140 pmol/L (20 mU/L), depending on the assay used. If doubt remains as to whether the subject is insulin-resistant, it may be necessary to check insulin and glucose concentrations after observed insulin administration or a hyperinsulinaemic clamp may be undertaken (see below).

If insulin-binding antibodies are suspected, for example in a subject who has received non-human insulins, free insulin can be assayed by immunoassay. If the free insulin concentration is much lower than total insulin concentration, insulin-binding antibodies are likely. Subjects who have little glycaemic response to high concentrations of endogenous insulin, but who are sensitive to exogenous insulin, may have an abnormal immunoreactive insulin (see p. 279 and Table 15.2).

If a subject without insulin-binding antibodies fails to respond to intravenous insulin, an insulin receptor or post-receptor problem should be suspected and a careful family study should be undertaken. Family members should be screened for insulin resistance by measuring fasting plasma glucose and insulin concentrations. The binding of insulin to the patient's white blood cells or adipose tissue may be measured in order to indicate whether the patient has normal insulin receptor numbers and whether the receptors show normal avidity for insulin.

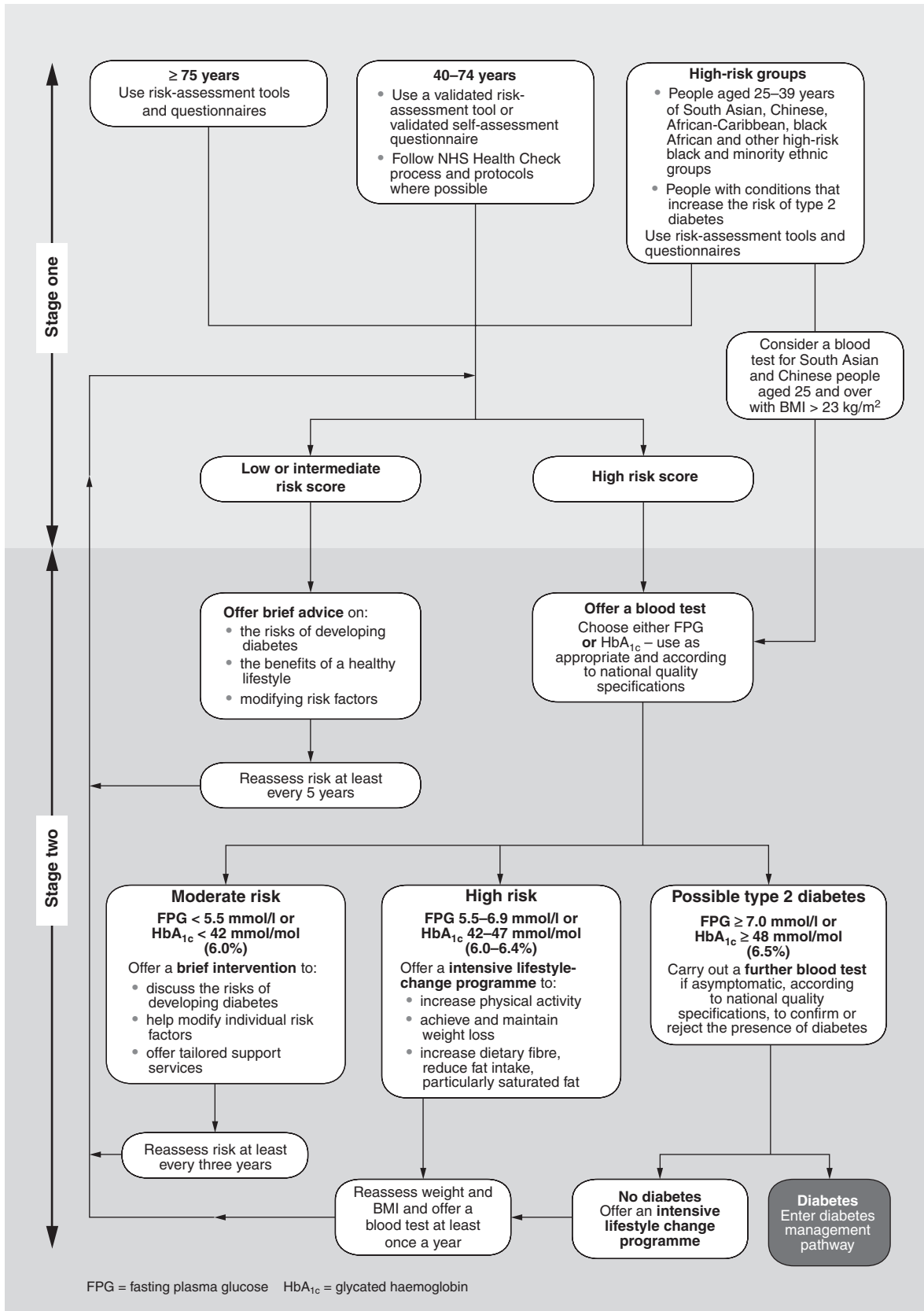


FIGURE 15.6 ■ Flowchart for the identification and management of type 2 diabetes mellitus. (From NICE public health guidance PH 38: Preventing type 2 diabetes: risk identification and interventions for individuals at high risk (2012): <http://guidance.nice.org.uk/PH38>, with permission from the National Institute for Health and Care Excellence).

Glucose transporter function can be investigated by incubating cells of interest (e.g. leukocytes, monocytes, adipocytes) with a non-metabolisable glucose analogue such as 2-deoxyglucose. The cellular content of the glucose analogue after a given time provides a measure of glucose transporter function. Glucose transporter number may be estimated by the use of specific antibodies linked to an imaging system (fluorescence or radioactivity based).

Molecular biological techniques can be used to determine DNA sequences coding for insulin, the insulin receptor and glucose transporter molecules, as appropriate, when the site of the defect has been identified.

Research investigations

Tests commonly used in diabetes research include the hyperinsulinaemic glucose clamp and the intravenous glucose tolerance test (IVGTT), as well as a variety of measures of β -Cell function.

Hyperinsulinaemic clamps. The hyperinsulinaemic clamp is the reference measure of insulin resistance. In the euglycaemic variant of the test, insulin is infused into a peripheral vein to increase the plasma insulin concentration to a target range (e.g. 250 pmol/m² per min or 35 mU/m² per min increases the insulin to around 60 mU/L). Without further intervention, the hyperinsulinaemia would induce hypoglycaemia. However, in the clamp procedure, the plasma glucose concentration is measured every 3–5 min and glucose is infused peripherally to maintain glucose concentrations within the desired range. When a steady state has been reached (usually 90–120 min), the rate of exogenous glucose infusion needed to maintain the glucose concentration is an index of the glucose clearance rate and of the subject's insulin sensitivity. Many variations on the clamp methodology are described, including use of stable isotope tracers, which permit better description of the physiological abnormalities in glucose turnover in response to insulin.

Intravenous glucose tolerance testing. The IVGTT is an alternative test of glucose tolerance in which glucose (typically 20 g/m² of body surface area) is injected as a bolus. This test has better reproducibility than the OGTT but is even less physiological. An IVGTT is a powerful stimulus to insulin secretion. It has been used in research studies to screen high-risk non-diabetic subjects for early defects of glucose tolerance and insulin response, with some success in predicting those that will subsequently progress to overt diabetes.

Measurement of β -Cell function. These tests are primarily used to determine the relative contribution of insulin deficiency and insulin resistance to glucose intolerance. There is no consensus as to which should be used in type 2 diabetes mellitus. The gold standard for research purposes is the hyperglycaemic clamp, where glucose is clamped at high concentrations and insulin secretion measured. If it is clinically important to document the severity of insulin deficiency, this is best done by measuring plasma C-peptide concentration. The DCCT trial used a basal concentration

for C-peptide of <0.2 nmol/L and a concentration of <0.5 nmol/L measured 6 min after giving 1 mg glucagon intravenously to distinguish type 1 from type 2 diabetes. Other tests of β -Cell function used for research purposes include the arginine stimulation test and mathematical models based on measures of insulin or C-peptide during oral or intravenous glucose tolerance tests.

Elevated fasting plasma proinsulin concentrations indicate subjects with abnormal β -Cell function, even if glucose tolerance is normal.

Homeostasis model assessment. The *homeostasis model assessment* (HOMA) is a mathematical model that uses fasting glucose and insulin concentrations to derive estimates of both insulin resistance and β -Cell function. It is useful for epidemiological studies where more sophisticated testing is not practical, but interpretation can be difficult especially in the context of treatment with insulin or oral drugs that stimulate insulin secretion.

CONCLUSION

There are two major types of diabetes: type 1 and type 2. Both are characterized by hyperglycaemia, but in type 1, this is a consequence of absolute insulin deficiency as a result of β -Cell destruction, whereas in type 2 the deficiency is relative, with most patients being resistant to the actions of insulin, although their production of insulin tends to decrease with time. Both conditions are associated with short-term and long-term complications. Biochemical investigations are essential for both the diagnosis and management of the conditions.

ACKNOWLEDGEMENT

The authors wish to acknowledge Dr Simon Coppack and Dr Victor Lawrence, who wrote this chapter in the second edition of this book.

Further reading

Bilous R, Donnelly R. *Handbook of diabetes*. 4th ed. Chichester: Wiley; 2010.

A good general account of diabetes.

Garber AJ, (ed.). *Type 2 diabetes*. In: *Med Clin North Am* 2004;88:787–1128.

This issue contains articles on the metabolic syndrome, and the pathogenesis, management and complications of type 2 diabetes.

Holt RIG, Cockram CS, Flyvbjerg A et al. editors. In: *Textbook of diabetes*. 4th ed Oxford: Blackwell Science; 2011.

A comprehensive account of all aspects of diabetes.

Kahn SE, Porte Jr D. Beta cell dysfunction in type 2 diabetes: pathogenesis and genetic basis. In: Valle D Beaudet AL, Vogelstein B et al. editors. *The online metabolic and molecular bases of inherited disease*. New York: McGraw-Hill. <http://dx.doi.org/10.1036/ommbid.86>.

MacLaren NK, Kukreja A. Type 1 diabetes. In: Valle D, Beaudet AL, Vogelstein B et al. editors. *The online metabolic and molecular bases of inherited disease*. New York: McGraw-Hill. <http://dx.doi.org/10.1036/ommbid.88>

Taylor SI. Insulin action, insulin resistance and type 2 diabetes mellitus. In: Scriver CR, Beaudet AL, Sly WS et al. editors. *The metabolic and molecular bases of inherited disease*. New York. <http://dx.doi.org/10.1036/ommbid.87>.

These three chapters provide detailed accounts of the pathogenesis of diabetes mellitus.

The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med* 1993;329:977–86.

A pivotal multicentre randomized study that provided firm evidence of the benefits of achieving and maintaining near-normal plasma glucose concentrations in patients with type 1 diabetes.

UK Prospective Diabetes Study (UKPDS) Group. Effect of intensive blood-glucose control with metformin on complications in

over-weight patients with type 2 diabetes (UKPDS 34). *Lancet* 1998a;352:854–65.

UK Prospective Diabetes Study (UKPDS) Group. Intensive blood glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet* 1998b;352:837–53.

A trial that clearly demonstrated the benefits of achieving and maintaining good glycaemic control in patients with type 2 diabetes.

The clinical management of diabetes mellitus

Ian W. Seetho • John P.H. Wilding

CHAPTER OUTLINE

INTRODUCTION 305

GENERAL ASPECTS OF MANAGEMENT 305

Nutrition 306

Exercise 306

Smoking cessation 307

Education about diabetes 307

Pharmacological management of cardiovascular risk 307

GLUCOSE-LOWERING THERAPY IN DIABETES 309

Background 309

Insulin use in type 1 diabetes 310

Glycaemic management in type 2 diabetes 313

Metformin 313

Sulfonylureas (and related insulin secretagogues) 314

Meglitinides 315

Peroxisome proliferator activator γ analogues 315

Glucagon-like peptide 1 analogues 316

Dipeptidyl peptidase IV inhibitors 317

Alpha-glucosidase inhibitors 317

Insulin use in type 2 diabetes 317

Bariatric surgery 317

Pancreatic transplantation 318

Islet cell transplantation 318

Immunotherapy for type 1 diabetes 318

OBSTACLES TO ACHIEVING GLYCAEMIC CONTROL 318

Intensive control 319

Hypoglycaemia 319

Intercurrent illness, 'sick day rules' and stress 321

CHRONIC COMPLICATIONS OF DIABETES 322

Nephropathy 322

Neuropathy 323

The feet in diabetes 324

Eye disease 325

Other complications 326

EMERGENCIES IN DIABETES 326

Diabetic ketoacidosis 326

Hyperosmolar hyperglycaemic state 329

Other metabolic acidoses 330

Alcoholic ketoacidosis 330

MANAGEMENT OF DIABETES IN THE HOSPITAL SETTING 330

PREGNANCY 331

CONCLUSION 332

INTRODUCTION

The rise in prevalence of diabetes presents some of the greatest medical and economic challenges of the 21st century. Despite advances in our understanding and management of diabetes mellitus, it continues to exact a huge toll both in terms of morbidity and mortality.

In this chapter, we consider dietary, lifestyle and macrovascular risk optimization measures for people with diabetes. Glucose-lowering treatment is discussed, followed by obstacles patients face in order to achieve their glycaemic goals, particularly hypoglycaemia. Later

sections discuss the management of chronic and acute complications of diabetes, and its management in particular circumstances, for example in critical illness and during pregnancy.

GENERAL ASPECTS OF MANAGEMENT

The management of patients with diabetes should involve a multidisciplinary team composed of diabetologists, diabetes specialist nurses, dietitians, podiatrists, the patient's general practitioner and other specialists if specific complications are present.

In addition to therapies that specifically target glycaemic control, nutritional management, lifestyle modifications such as adequate physical exercise and smoking cessation, and structured patient educational programmes are key components in the management of patients with diabetes.

Nutrition

Patients with diabetes should follow a healthy diet and make appropriate nutritional choices to reduce cardiovascular risk and improve glycaemic control. A dietitian with specialist knowledge in diabetes should take the lead role in providing nutritional care. Dietary advice should be tailored specifically for each patient, considering their personal and cultural preferences, beliefs, lifestyle and the changes that the patient is willing and able to make.

The principles of diet include eating regular meals with a reduction in simple carbohydrate intake with an increase in complex carbohydrates, a high fibre content and an energy intake to achieve or maintain an ideal body weight, or in those who are overweight or obese, to achieve and sustain modest weight loss. The key aim is to balance energy input and expenditure and ensure a good quality diet.

Contrary to the general perception that a diet in diabetes needs to be unique or special, there is no special 'diet for diabetes'. People with diabetes, except perhaps in specific circumstances such as renal failure, should eat the same healthy diet as recommended to those who do not have diabetes. Thus, no more than 30% of total energy intake should be in the form of fat, and the ratio of unsaturated to saturated fats should be increased with saturated fat intake ideally making up <7% of total calories and *trans* fatty acid intake minimized. Replacement of saturated fats with mono- or polyunsaturated fats aids lowering of cholesterol.

Carbohydrate (preferably in the form of slowly absorbed complex carbohydrate with a low 'glycaemic index' rather than simple sugars) should provide approximately 55% of total energy intake with protein providing the remaining 15%. Foods that improve glycaemic control and might reduce cardiovascular risk include whole grains and green vegetables. Conversely fried foods, high glycaemic index diets and red and processed meats may increase cardiovascular risk.

Carbohydrate intake can significantly affect postprandial blood glucose concentrations. In type 1 diabetes, matching insulin doses to the amount of carbohydrate consumed (carbohydrate counting) is effective in improving glycaemic control. Patients who are on fixed insulin doses or biphasic insulin regimens should be advised to aim for consistency in carbohydrate and glycaemic index dietary intake.

In patients treated with diet only or diet and oral hypoglycaemics, total energy intake in terms of carbohydrate consumption should be monitored as this is important in achieving glycaemic control. Strategies employed may include carbohydrate counting, exchanges or food portions based on past experience. However, as each individual is different, there is no fixed recommended ideal macronutrient composition of carbohydrate, fat and protein in the

diet. Nevertheless, some studies have shown improved body weight and glycaemic control with low carbohydrate diets through a reduction in energy intake over the short term.

Sodium intake should not generally exceed 6g/day and plentiful fruit and vegetables ('five portions a day' being the current mantra) should be consumed, mindful of the fact that some fruits (e.g. grapes) may contain large amounts of sugar and should not be taken in excess. A total daily dietary fibre intake of 40g is ideal but rarely achieved in practice. Except in specific and unusual circumstances, it is a general principle that the pharmacological treatment of diabetes should be tailored to the dietary intake and lifestyle of the individual and not vice versa. Fructose reduces postprandial glycaemia when it is used as a replacement for sucrose or starch, but it may have adverse metabolic effects including causing dyslipidaemia and insulin resistance. Non-nutritive sweeteners are safe when consumed within recommended daily limits.

It is currently recommended that patients with new onset type 2 diabetes should be managed with dietary and lifestyle modification with consideration for early introduction of metformin (discussed later). The degree of adherence to a particular diet will affect the outcome and a diet that is both palatable and acceptable is more likely to be followed.

Exercise

Patients with diabetes should be encouraged to participate in regular exercise of the kind that would benefit the rest of the population. Regular low-intensity exercise, such as brisk walking (so as to be slightly breathless while talking at the same time), swimming or cycling for 30 min three to five times a week should be a realistic goal for many individuals. Exercise improves glucose disposal (by increasing the expression of type 4 glucose transporters, GLUT4) in skeletal muscle, among other mechanisms), prevents progression from impaired glucose tolerance to type 2 diabetes (by ~50%), increases basal metabolic rate and has convincingly been demonstrated to reduce the risk of cardiovascular events.

Some patients with cardiovascular disease require detailed assessment (e.g. electrocardiogram, review of medications) before taking part in exercise programmes, but the few absolute contraindications to exercise in patients with diabetes (e.g. tight aortic valvular stenosis) apply equally to individuals without diabetes.

In insulin-treated patients, exercise may be associated with either an increased frequency of hypoglycaemia or hyperglycaemia as blood glucose concentrations are influenced by the timing, type and quantity of insulin, carbohydrate intake and intensity and duration of physical activity. Careful monitoring of blood glucose concentrations is necessary to guide the adjustment of insulin therapy and carbohydrate intake, and the strategy adopted may depend on whether the exercise is planned or unplanned.

Aerobic exercise improves glycaemic control and lowers low density lipoprotein (LDL) cholesterol in patients with type 2 diabetes, but has little effect on other lipid parameters. Resistance training has effects on both

glycaemia and cardiovascular risk factors such as systolic blood pressure, fat mass and insulin resistance.

Studies have shown that it is safe for individuals with type 2 diabetes who are treated by diet alone or in conjunction with oral hypoglycaemic agents to exercise in both the fasting and post-meal states with the most beneficial effects on blood glucose concentrations observed postprandially. As with dietary modification, exercise programmes must be individually designed in order to encourage compliance.

Smoking cessation

Smoking is hazardous to all, but is particularly so in people with diabetes. The macrovascular risks of smoking are compounded with those of diabetes itself. Smoking is associated with increased mortality and predisposes to microvascular complications of diabetes such as retinopathy and nephropathy. Smoking is potentially diabetogenic: it is a risk factor for the development of type 2 diabetes, and can influence glycaemic control by impairing insulin sensitivity and glucose tolerance.

Smoking is at least as prevalent in patients with diabetes as in the general population and young women in particular may perceive the habit as an adjunct to weight management.

Many approaches to the supportive management of patients have been described. These include smoking-cessation clinics and nicotine replacement therapy. Such forms of therapy apply equally to people with and without diabetes.

Education about diabetes

Structured educational programmes have been developed in Europe and North America. In the UK, a five-day intensive educational course called the Diet Adjustment For Normal Eating (DAFNE) programme provides opportunities for support and learning for patients with type 1 diabetes. This course teaches those with type 1 diabetes how to adjust their insulin doses according to carbohydrate portion intake, and allows participants to learn about aspects of lifestyle and management related to their diabetes.

For patients with newly diagnosed type 2 diabetes, the Diabetes Education and Self Management for Ongoing and Newly Diagnosed (DESMOND) programme offers support to patients in identifying their health risks and enables goal-setting, developing confidence in self-management of their condition.

Pharmacological management of cardiovascular risk

Despite the fact that patients with diabetes are uniquely susceptible to microvascular complications, around 80% of patients with type 2 diabetes will die of macrovascular disease. Ischaemic heart disease, stroke and peripheral vascular disease are some 2–4 times more common in patients with diabetes than in the non-diabetic population. Many available risk tables do not even include powerful diabetes-related risk factors such as microalbuminuria.

Type 2 diabetes is associated with dyslipidaemia. There are elevated concentrations of small, dense LDL particles that are more prone to oxidation, which increases their ability to damage the vascular endothelium. Reduced high-density lipoprotein (HDL) concentrations result in reduced protection against atheroma: increased triglyceride-rich lipoprotein concentrations predispose to atheroma formation.

The risk of mortality from cardiovascular disease is higher in patients with diabetes and increases further with additional cardiovascular risk factors such as hypercholesterolaemia, hypertension and smoking. Diabetes is so powerful a risk factor that patients should, in most cases, be treated as if they had already suffered a cardiovascular event, rather than have their treatment assessed on the basis of primary prevention studies.

Apart from recommendations regarding the dietary and other lifestyle measures detailed above, there is a degree of uncertainty about which patients should receive preventative treatment with agents such as aspirin, lipid-lowering therapy or angiotensin-converting-enzyme (ACE) inhibitors, and what targets to pursue at any specific stage of the disease. Guidelines have been prepared by various organizations in countries with populations of differing genetic background and socioeconomic structures. None can obviate the need for individualized risk–benefit assessment, and it is important to involve patients in setting their own treatment goals, and management plans.

It is likely that future guidelines will recommend a lifetime risk calculator that uses a range of risk factors to calculate the absolute risk of an event over the rest of the patient's lifetime. This has the effect of promoting more aggressive intervention to modify risk factors from an early age. Risk calculators can also be used to help patients to understand their personal risk and how changes can affect it, in order to improve compliance.

Aspirin

Aspirin (75 mg daily) is recommended for secondary prevention of cardiovascular disease in patients with diabetes. It is effective in reducing cardiovascular morbidity and mortality in high-risk patients with previous cardiovascular disease. The risk of adverse effects associated with this dose of aspirin depends on many factors, including the age of the patient, intake of other drugs (corticosteroids and non-steroidal anti-inflammatory drugs), presence of other disease (e.g. untreated proliferative retinopathy, uncontrolled hypertension, asthma) and medical treatment (e.g. forthcoming surgery). The risks of major gastrointestinal haemorrhage appear to be increased by a factor of 1.5–2 with approximately one additional event per 500 patient years under trial conditions (approximately 0.2% per patient per year). Intracranial haemorrhages are increased by one event per 3–500 patient years.

The use of aspirin for secondary prevention of macrovascular disease in all patients in whom it is not contraindicated is well established. However, in primary prevention in patients with diabetes with no prior cardiovascular events, the use of aspirin has been more contentious.

The current Joint British Societies guidance (JBS2) recommends aspirin 75 mg daily in patients with diabetes aged >50 years, those who are younger but have had the disease for more than 10 years, and those who are already receiving treatment for hypertension, with controlled blood pressure <150 mmHg systolic and <90 mmHg diastolic. The UK National Institute for Health and Care Excellence (NICE) guidelines have stated that low-dose aspirin should be offered to patients aged >50 years if blood pressure is below 145/90 mmHg and to those aged <50 years who have significant cardiovascular risk (features of the metabolic syndrome, strong early family history of cardiovascular disease, smoking, hypertension, existing cardiovascular disease or microalbuminuria).

However, a more recent meta-analysis found no significant beneficial effects of using aspirin for primary prevention of cardiovascular disease in diabetes, and the Prevention of Progression of Arterial Disease and Diabetes (POPADAD) study demonstrated that aspirin did not significantly reduce the risk of major cardiovascular events even in patients with asymptomatic peripheral arterial disease. At present, there is little evidence to support the use of aspirin in patients aged <30 years and, in those <16 years, aspirin should be avoided because of an increased risk of Reye syndrome.

Lipid-lowering agents

Effective management of dyslipidaemia reduces the risk of cardiovascular disease; plasma lipid concentrations should be monitored at least annually in patients with diabetes. Most people in the developed world, with or without diabetes, would benefit from a less atherogenic lipid profile. Lifestyle modifications such as dietary restriction of saturated fat, cholesterol and *trans*-unsaturated fat, increased intake of ω -3 fatty acids, viscous fibre and plant sterols, achievement and maintenance of an ideal body weight and taking adequate exercise are recommended, whether or not further treatment is required. Secondary causes of dyslipidaemia such as alcohol excess, hypothyroidism or liver disease should be considered in all patients prior to initiating any lipid-lowering therapy. For patients with diabetes, poor glycaemic control (denoting inadequate insulin action) may also cause excess very low density lipoprotein (VLDL) production and hypertriglyceridaemia. Treatment with metformin, pioglitazone and insulin have all been shown to improve this, either via increasing insulin action or by other specific mechanisms (e.g. by reducing the flux of non-esterified fatty acids to the liver in the case of pioglitazone).

Many primary and secondary prevention trials have convincingly demonstrated the benefits of lipid-lowering therapy in reducing macrovascular risk in subjects with and without diabetes. Many of these studies have involved the use of 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) reductase inhibitors (statins), and have shown that lowering LDL-cholesterol is approximately linearly related to risk reduction down to concentrations of 2.2 mmol/L or even below.

The American Diabetes Association (ADA) currently recommends statin therapy, in addition to lifestyle factors, as secondary prevention for patients with diabetes who have

cardiovascular disease, with a goal for LDL-cholesterol of <1.8 mmol/L. Treatment is recommended for primary prevention in all patients with diabetes over the age of 40, and in younger people who have additional cardiovascular risk factors, with an LDL-cholesterol target of <2.6 mmol/L. Statins should be used with caution, and only on specialist advice, in children and are contraindicated in pregnant women.

The typical diabetic dyslipidaemia comprises raised triglycerides, intermediate density lipoprotein (IDL) and small, dense LDL concentrations together with low HDL concentrations. These abnormalities are particularly atherogenic, with the result that the significance of dyslipidaemia in diabetes is underestimated by routine measurements of total lipid subfractions. Fibrin acid derivatives have a role in the management of the raised triglycerides in diabetic dyslipidaemia. The ADA currently recommends targets for triglycerides of <1.7 mmol/L, and for HDL-cholesterol of >1.0 mmol/L in men and >1.3 mmol/L in women. Both the JBS2 and NICE guidelines (NICE Clinical Guideline 66) recommend a target total cholesterol <4.0 mmol/L and LDL <2.0 mmol/L. Alternatively, JBS2 recommends a 25% reduction in total cholesterol with a 30% reduction in LDL-cholesterol, if this achieves lower absolute concentrations. Both guidelines recommend statins for those with diabetes who are aged >40 years and for those aged 18–39 years with diabetes complications (retinopathy and nephropathy), poor glycaemic control, hypertension, features of the metabolic syndrome or family history of premature CVD in a first degree relative.

Combination lipid-lowering therapy, to achieve additional lowering of LDL, may be considered in order to reach desired targets if these are not reached with maximum tolerated statin doses. To date, however, there is little evidence to show that combination therapy to lower LDL produces significant cardiovascular risk reduction over and above statin therapy alone. We would advocate a strategy of: (1) assessing overall risk and choosing targets for LDL, triglyceride and HDL; (2) excluding secondary causes of hyperlipidaemia and optimizing glycaemic control; (3) optimizing lifestyle options, and (4) targeting LDL-cholesterol with a statin. These drugs may in addition produce minor favourable effects on HDL-cholesterol and triglycerides. It should be noted that therapy targeting HDL or triglycerides does not have the evidence base of statin treatment.

Severe hypertriglyceridaemia requires lifestyle and pharmacological therapy. A fibrate or ω -3 fatty acid-rich marine oils may be used to reduce the risk of pancreatitis associated with high triglycerides. Any statin–fibrate combination requires care because of the potentially increased risk of rhabdomyolysis, and monitoring of plasma creatine kinase activity may be appropriate.

Hypertension

Assiduous control of blood pressure (BP) has been shown to reduce major cardiovascular events and minimizes progression of diabetes complications. In the UK Prospective Diabetes Study (UKPDS), intensive treatment of hypertension to a median of 144/82 mmHg was compared with conventional BP treatment (median BP in the control

group was 154/87 mmHg). Intensive treatment resulted in reductions in multiple diabetes-related endpoints, including death. The majority of subjects in the tight control arm of the study required three or more agents in order to achieve this. The ADA has adopted an even more aggressive target of 130/80 mmHg on the basis of epidemiological risk studies showing an inflexion in the relationship between BP in patients with diabetes and cardiovascular risk at 115/75 mmHg.

Apart from more rigorous treatment goals, management of hypertension in patients with diabetes is similar to that in patients without the condition. Initially, lifestyle changes are recommended including weight loss, exercise and dietary modifications, such as a reduction in sodium and alcohol intake and increased consumption of fruit and vegetables. However, if blood pressure remains inadequately controlled despite lifestyle modifications, pharmacological therapy may be commenced.

Some patients will have comorbid conditions that will direct the choice of drug therapy. The presence of microalbuminuria, proteinuria, mild to moderate renal impairment, diabetic retinopathy, ischaemic heart disease, stroke and left ventricular systolic dysfunction would, for example, indicate a specific role for an angiotensin-converting-enzyme (ACE) inhibitor or angiotensin-II receptor antagonist (blocker) (ARB). Given the potentially beneficial effects of ACE inhibitors and ARBs on glycaemia and in the prevention of diabetic complications such as diabetic nephropathy (see below), they are generally considered as first-line agents in patients with diabetes if lifestyle changes do not improve blood pressure.

β -Blockers, diuretics, and calcium channel blockers have all been shown to reduce cardiovascular events in patients with diabetes, and all may be used in combination with an ACE inhibitor or ARB. Most patients will require multiple agents to manage their hypertension. Generally initial treatment will be with either an ACE inhibitor or ARB, and if required, a calcium channel blocker, then a thiazide diuretic, β -blocker, α -blocker or potassium-sparing diuretic. It is important to monitor renal glomerular function and plasma potassium concentration whenever ACE inhibitors, ARBs or diuretics are used. Thiazides are effective, but unwanted effects including hyperglycaemia, erectile dysfunction, gout and dyslipidaemia have reduced their use as first line agents. β -Blockers are specifically indicated in some situations, for example stable heart failure and ischaemic heart disease, and were found to be safe and effective in the UKPDS. However, they may promote weight gain and, in some patients with impaired awareness, hypoglycaemia may be a problem. In the Anglo-Scandinavian Cardiac Outcome Trial (ASCOT), treatment with an atenolol-thiazide-based regimen was associated with 30% greater incidence of new type 2 diabetes than treatment with an amlodipine-perindopril-based regimen, and was less effective at reducing BP and cardiovascular events in patients with established diabetes.

In the past, there were concerns about the safety of the dihydropyridine group of voltage activated calcium channel blockers. However, results from the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) and ASCOT have been reassuring, and

it now seems reasonable to recommend amlodipine as a second-line agent in combination with an ACE inhibitor or ARB (although amlodipine and other calcium channel blockers inhibit the metabolism of simvastatin, which must not be co-prescribed at high dose).

α -Blockers appear effective and have favourable effects on insulin sensitivity and LDL-cholesterol concentrations. However, ALLHAT showed an apparent doubling in the incidence of heart failure (a secondary outcome measure) with an α -blocker compared with a thiazide diuretic. For this reason, α -blockers are generally reserved for use where treatment with at least two other classes of hypotensives in combination has failed to achieve the target BP.

Angiotensin-converting-enzyme inhibitors and angiotensin-II receptor antagonists

Angiotensin II (ATII) increases hepatic glucose production and decreases insulin sensitivity. The use of ACE inhibitors or ARBs consistently increases insulin sensitivity by some 15–20% in pharmacological studies. The Heart Outcomes Prevention Evaluation (HOPE) study was terminated early after interim analysis demonstrated that 4.5 years of treatment with the ACE inhibitor ramipril, 10 mg daily, reduced major vascular events and death in subjects with preserved left ventricular function, and either pre-existing vascular disease or those with diabetes and an additional vascular risk factor. The latter group also had a reduced risk (relative risk 0.84) for the development of diabetic complications. In an extension to the study to a total of 7.2 years follow-up, there was a 31% relative risk reduction for new diagnosis of type 2 diabetes.

Results of other studies and meta-analyses have, in general, not shown such large benefits in patients who already have diabetes (unless with other clearly established indications), although they have generally been restricted to relatively small subgroup analyses. While it is not clear whether all (non-pregnant) patients with diabetes should be offered these agents, ACE or AT-II inhibition is clearly indicated for subjects with diabetes and hypertension, microalbuminuria, proteinuria, mild to moderate renal impairment, diabetic retinopathy, ischaemic heart disease, stroke or left ventricular systolic dysfunction. Caution must be exercised with the use of these agents when renal artery stenosis is present (e.g. in patients with impaired renal function and absent distal pulses or a renal bruit) and in subjects with hyporeninaemic hypoadosteronism (type 4 renal tubular acidosis), in whom dangerous hyperkalaemia may result.

GLUCOSE-LOWERING THERAPY IN DIABETES

Background

Trials in both type 1 Diabetes Control and Complications Trial (DCCT) and type 2 (UKPDS) diabetes have confirmed that microvascular, and, to a lesser extent, macrovascular, complications of diabetes may be delayed or even prevented by tight glycaemic control. The UKPDS (1998) trial in type 2 diabetes found that a 10 mmol/mol

(0.9%) reduction in HbA_{1c} from 63 to 53 mmol/mol (7.9–7.0%) over a median follow-up of ten years from diagnosis led to a reduction of 12% for any diabetes-related endpoint, 25% for microvascular endpoints and 24% for cataract extraction, and of 21% for retinopathy and 33% for albuminuria at 12 years. The improvements in macrovascular disease were less convincing and of borderline statistical significance.

The DCCT (1993) trial in type 1 diabetes showed a risk reduction of 26–63% in microvascular disease with intensive insulin therapy, but at the cost of a three-fold elevation in the incidence of severe hypoglycaemia (fits, unconsciousness) and some additional weight gain compared with the control arm of the study. However, the use of more modern insulin analogues has been associated with lower incidence of hypoglycaemia.

Insulin use in type 1 diabetes

Patients who have been diagnosed with type 1 diabetes will require insulin therapy along with general measures such as dietary and lifestyle changes. Although it is often considered that patients with type 1 diabetes have no useful islet cell function, this is not entirely true. In fact, β -Cell function is approximately 10% of normal at disease presentation, and may double after the initiation of insulin therapy, resulting in a reduced requirement for exogenous insulin and stabilization of glucose metabolism ('honeymoon effect'). This effect does not last and, in the majority of patients, insulin secretion declines again over the next 2–5 years.

The 'honeymoon effect' may reflect, in part, the amelioration of glucotoxicity or lipotoxicity (see Chapter 15) on the reduced numbers of metabolically stressed β -Cells. After initiation of insulin therapy, the toxic effects of hyperglycaemia are removed, thereby relieving this metabolic stress. In some studies, as many as 10% of patients with newly diagnosed type 1 diabetes have been maintained off insulin altogether for the first year, but this is clearly not recommended in routine clinical practice because of the risks of ketoacidosis.

In a subset (~15%) of patients, there is retention of some detectable islet function, as assessed by C-peptide secretion, for at least 40 years. These individuals also appear to have reduced susceptibility to severe diabetic ketoacidosis (DKA), less risk of retinopathy, a more robust glucagon response to hypoglycaemia and lower average insulin requirements.

A number of trials have attempted to augment residual insulin secretion with immunosuppressive drugs (e.g. steroids, azathioprine, antithymocyte globulin and ciclosporin) or by plasmapheresis, and these measures have been met with some success but, on current evidence, their benefits have not been clearly outweighed by their risks. Studies using strict metabolic control in order to mitigate any effects of lipotoxicity or glucotoxicity on the β -Cell (e.g. with continuous subcutaneous insulin infusion, CSII) have been able to demonstrate transiently preserved insulin secretion, but only for a period of a few months, with rapid regression on discontinuation of strict metabolic control.

Regular insulin

Regular human insulin (e.g. Actrapid[®], Humulin S[®]) has an onset of activity within 30–60 min after subcutaneous injection, a variable peak concentration at around 2–4 h, and a duration of action of 5–8 h, contrasting with its duration of action of only a few minutes when administered intravenously. This difference may be explained by the tendency for hexamerization in solution.

The hexameric form of regular insulin has to dissociate into dimers or monomers in order to be absorbed into the bloodstream. This is rate-limiting for its absorption and influences the pharmacokinetic profile. As such, in order to treat prandial glucose peaks, regular insulins have to be injected about 30 min before each meal so that the onset of action coincides with the rise in prandial glucose concentrations.

Insulin analogues

More recent developments in insulin technology have focused on the design and production of insulin analogues whose pharmacokinetic profiles are determined by modifications to non-receptor-binding regions of the insulin molecule.

Insulin aspart (Novorapid[®]) is homologous with regular human insulin with the exception of a single substitution of aspartate for proline in position B28 of the B chain: it is produced by recombinant DNA technology in *Saccharomyces cerevisiae* (baker's yeast). Insulin lispro (Humalog[®]) is identical to normal human insulin except that the positions of the amino acids proline and lysine at B28 and B29 are reversed. In insulin glulisine (Apidra[®]), asparagine is replaced by lysine at position B3, with glutamic acid replacing lysine at B29. As a result of these changes in amino acid sequence, the rapid insulin analogues (aspart, lispro and glulisine) have a lower tendency to form hexamers, permitting more rapid absorption and a faster onset of action (<15 min), enabling injection immediately before, or just after, eating. There is a sharper peak, more closely resembling first-phase insulin secretion in the non-diabetes state, and a shorter duration of action (3–5 h), which may reduce the incidence of hypoglycaemia before the next meal is ingested.

Intermediate-acting insulin

Isophane (NPH, Neutral protamine Hagedorn) is an intermediate-acting insulin that is a combination of soluble insulin and protamine zinc. Protamine delays the onset of action and prolongs the effects of the soluble insulin. The onset of action is within two hours, with a duration of action of 10–20 h. Most patients require two subcutaneous injections daily to provide adequate glycaemic control. Isophane has fallen out of favour because of increased risks of nocturnal hypoglycaemia and fasting hyperglycaemia.

Premixed insulin analogues

The most commonly prescribed premixed insulins are a mixture of quick acting analogue and an insulin analogue protamine complex (e.g. Novomix 30[®] is composed of

insulin aspart-protamine complex and insulin aspart in a ratio of 70:30). These are usually given as a twice-daily regimen.

Long-acting insulin analogues

Longer-acting insulin analogues (insulin glargine and insulin detemir) are produced by genetic engineering. The onset of action is within two hours and they have a longer duration of action of up to 24h. These insulins provide a steady basal insulin profile with minimal peak action and are injected subcutaneously once daily.

In insulin glargine, asparagine is replaced by glycine at position 21 of the α chain of the insulin molecule and the carboxyl terminal of the β chain is changed with the addition of arginine. This insulin is slowly released into the circulation because of its acidic pH, which causes it to precipitate when injected subcutaneously. In insulin detemir, tyrosine at position B30 is lost with acylation at B29 that results in a fatty acid side chain (tetradecanoic acid) becoming linked with lysine. With these alterations, insulin detemir forms hexamers with slower dissociation and a longer duration of action.

Insulin regimens

Most patients with type 1 diabetes use an insulin replacement regimen that comprises multiple-dose subcutaneous injections (three to four injections per day) consisting of basal and prandial insulin. Insulin is injected subcutaneously. Short-acting insulin (regular soluble or analogue) acts to control prandial blood glucose concentrations; long-acting insulin provides a constant background basal insulin concentration. This combination of short-acting and long-acting basal insulin is called the basal bolus regimen or multiple daily injection therapy. Most people take their long-acting insulin at bedtime in order to provide optimal glycaemic control overnight. Some patients learn how to adjust insulin doses according to carbohydrate intake (in the UK often through attendance at a DAFNE (dose adjustment for normal eating) course).

An alternative is a twice daily insulin regimen consisting of a premixed or biphasic mixture of longer- and shorter-acting insulins in ratios of 75:25, 70:30 or 50:50. Such regimens are commonly used in type 2 diabetes and in patients with type 1 diabetes before they progress to a basal bolus regimen. Some patients with type 1 diabetes prefer twice daily insulin to the basal bolus regimen, particularly those who dislike frequent injections, but it is not suited to those whose diet and activities tend to differ on a day-to-day basis, as there is less scope for dose adjustment in relation to carbohydrate intake, with attendant risks of hypoglycaemia. The longer-acting component may not provide effective glycaemic control overnight because of a waning of the effect of the previous evening's dose, a decrease in insulin sensitivity in the early hours of the morning, and increases in growth hormone and cortisol concentrations, all of which will increase the risk of early morning hyperglycaemia (the 'dawn' phenomenon). Premixed insulin preparations are, therefore, not recommended for long-term treatment

in type 1 diabetes and most patients are eventually recommended to convert to the basal bolus regimen for optimal glycaemic control. The relative merits and disadvantages of these two regimens, and a comparison with continuous subcutaneous insulin infusion, are outlined in Table 16.1.

Continuous subcutaneous insulin infusion

Continuous subcutaneous insulin infusion pumps ('insulin pumps') are portable devices that deliver a continuous subcutaneous infusion of short-acting insulin at a set rate. The pump is battery-driven and comprises a reservoir containing about 3 mL of insulin in a syringe. A syringe driver delivers insulin via a cannula that is inserted subcutaneously into the abdominal wall. The cannula is re-sited every 1–2 days to avoid skin infections. In addition to the continuous basal infusion, the patient triggers the delivery of an additional dose at mealtimes to provide prandial cover.

The advantages of a pump include having a basal insulin rate that is personalized to each individual, allowing more flexibility with regards to mealtimes as the basal rate can be supplemented with a corrective mealtime insulin dose according to the carbohydrate consumed. Studies have shown improved glycaemic control in pump users. However, in the event of pump failure, there is the risk of diabetes ketoacidosis occurring.

Clinical indications for pump therapy include poor glycaemic control despite high levels of care, disabling hypoglycaemia on multiple daily injection regimens, recurrent unpredictable hypoglycaemia, especially in patients with severe dawn phenomenon and pregnant women without acceptable glycaemic control with subcutaneous insulin. Before each patient is commenced on the insulin pump, they must attend a comprehensive educational pump course. They must be sufficiently motivated to use the pump, with a commitment to regular blood glucose monitoring, attendance at pump clinics and adherence to carbohydrate counting so that they will be able to manage their pump. Close liaison with the diabetes pump team is essential for the success of this form of therapy.

It should be stressed that when deciding which regimen to use, each patient's specific circumstances will need to be carefully considered in light of different needs, goals and capabilities and the level of support available. An important factor is the risk of hypoglycaemia that comes with insulin therapy. There should be a balance between achieving optimal glycaemic control whilst at the same time ensuring each patient's safety, in terms of hypoglycaemia risk, when deciding on which insulin and regimen is to be used.

Insulin administration

The recommended subcutaneous injection sites include the abdomen, upper outer thighs, upper outer arms and buttocks. It is important to avoid injecting insulin intramuscularly. This commonly occurs in the limbs of slim individuals or children and can influence insulin absorption and hence blood glucose profiles.

TABLE 16.1 Comparison of different insulin regimens

	Twice daily mixed insulin	Basal-bolus regime	Constant subcutaneous insulin infusion (CSII)
Glycaemic control	Some patients achieve excellent results but, in general, the method is inferior to basal-bolus insulin regimens and most adult patients with type 1 diabetes requiring good control should probably use an alternative regimen	Potential for very good control in most individuals. Little evidence that it is systematically less efficacious than CSII	Potential for the best possible control, but training and considerable patient expertise required. May not be suitable for patients with totally erratic control. Expensive (both unit cost and disposables)
Flexibility of eating	Hypoglycaemia during mid morning a specific problem without snacking. Most suitable for those whose diet and lifestyles are similar from day-to-day	Basal insulin dose usually sufficient to avoid ketoacidosis if a meal and its associated insulin are missed. Additional correction doses may be given if pre-meal blood glucose is elevated. Additional dose may be given to cover unscheduled eating, e.g. a snack, ice-cream	Greatest possible flexibility
Hypoglycaemia risk	Particular risk if lunch delayed or small. Risk if dose is adjusted, e.g. to cover a large meal without considering the effect this will have on the longer-acting insulin in the mixture. The longer-acting insulin component used in the early evening rather than before bed risks morning hyperglycaemia (dawn phenomenon), and increases in dose may risk hypoglycaemia earlier in the night, as peak action (6–8 h) may occur after maximum growth hormone secretion in early deep sleep	Depends on doses used and degree of control aimed for. Some evidence that insulin analogues maintain/improve glycaemic control with reduction in hypoglycaemia	Depends on doses used and degree of control aimed for. User error may lead to hypoglycaemia
Number of injections	Lowest possible in type 1 diabetes (two per day)	More injections (4–5 depending on whether a single dose of longer-acting insulin achieves adequate 24 h cover)	No injections (abdominal cannula left in situ for 24–48 h) but the needle site may be irritating/painful. Infections at insertion site are unusual but very unpleasant
Dosage titration principles	Titrate morning dose to pre-evening meal blood glucose and evening dose to pre-breakfast blood glucose. Difficult to adapt to a specific meal or situation (e.g. exercise) as adjustment of each insulin independently is not possible (use of different mixes at different times theoretically possible but rarely done). May be particularly suitable for patients who prefer not to titrate insulin doses themselves and who have little day-to-day variation in lifestyle. Consider changing the mix if necessary to provide more or less soluble insulin based on blood glucose measurements 2 h after breakfast and evening meal. Basic regimen may be adapted, e.g. supplemented with soluble insulin at lunchtime	Titrate basal insulin to fasting (especially pre-breakfast) blood glucose concentrations and bolus insulin to 2 h postprandial results for each meal. Carbohydrate counting (e.g. the dose adjustment for normal eating (DAFNE) principle) may be applied, noting that greater doses may be required for the same effect in the morning than in the evening. Correction doses may be applied according to pre-meal blood glucose concentrations if prepared to test at least thrice daily. Basic regimen may be adapted, e.g. by using an insulin mixture with breakfast to avoid the need for a lunchtime injection. Type of each insulin may be adapted to need, e.g. Actrapid® vs NovoRapid®, depending on the interval between meals and blood glucose values by the time of the next meal after optimal titration of basal insulin	Complex programming is possible (e.g. different regimens for work days and weekends). Typically 50% of the total daily insulin is given as basal insulin at 1–2 units/h during the day and 0.5 units/h at night, increasing to coincide with the dawn phenomenon with the remaining 50% given as mealtime boluses depending on the timing, size, content and glycaemic index of the meal to be ingested. Basal rates may be varied according to other activities, e.g. exercise, illness
Other	Difficult to use when sick day rules apply: risk of ketoacidosis	Less risk of ketoacidosis as patient may titrate short-acting insulin to blood glucose taken 4–6 hourly when ill, noting that the total daily dose may need to be increased when ill	Risk of ketoacidosis associated with mechanical failure (e.g. pump stops, needle falls out, blocks) becoming less of a problem with advances in design and alarming of systems

As a result of the trophic action of insulin, an accumulation of fat, called lipohypertrophy, may occur at sites of repeated subcutaneous insulin injections. Lipohypertrophy can affect insulin absorption, causing problems such as hypoglycaemia, but can be avoided by rotating injection sites. Insulin injection sites should be inspected when patients attend for their diabetes review.

Glycaemic management in type 2 diabetes

When type 2 diabetes is diagnosed, lifestyle modification should be advised initially, and metformin commenced in combination with lifestyle if HbA_{1c} does not reach target after three months. If glycaemic control remains poor after 3–6 months, further oral agents, or a glucagon-like peptide 1 (GLP-1) receptor agonist or insulin may be added. Some patients with marked hyperglycaemia may be commenced on insulin at diagnosis if it is deemed that non-insulin agents will not provide the necessary control of blood glucose.

Metformin

Metformin belongs to a class of drugs called the biguanides that act to reduce insulin resistance, therefore improving insulin sensitivity. Metformin is the only biguanide currently available. It has an established role in the treatment of type 2 diabetes in combination with diet and lifestyle measures. The UKPDS study showed that obese patients treated with metformin had reduced myocardial infarction (by 6%), stroke (by 3%) and mortality (by 7%). These effects were independent of improved glycaemic control.

Metformin has the clear advantage over many other drugs of improving glycaemic control without causing weight gain. Metformin does not cause hypoglycaemia and may even lead to weight loss by suppressing appetite. As such, it is often prescribed as first line treatment in patients who are overweight. Several of the adverse features of the metabolic syndrome, including raised circulating concentrations of non-esterified fatty acids (NEFAs), triglycerides and plasminogen activating inhibitor-1 (PAI-1), are improved by metformin. Metformin has an effect on blood glucose that is additive to, and broadly equivalent to, that of sulfonylurea monotherapy. In type 2 diabetes, metformin can be used as monotherapy, or combined with insulin or with sulfonylureas, dipeptidyl peptidase IV (DPP-4) inhibitors, GLP-1 analogues or thiazolidinediones. In type 1 diabetes it is sometimes used together with insulin for obese adults, although this usage is currently unlicensed. Metformin may also be considered in patients who are not overweight.

Metformin has a useful role in diabetes prevention; a 31% relative reduction in the progression of diabetes was observed in the Diabetes Prevention Programme (DPP) placebo-controlled study, with greatest effect in younger and more obese subjects. Metformin also has a possible role in the treatment of polycystic ovary syndrome (PCOS), alone or in combination with drugs acting on sex steroid metabolism, particularly in those patients who are also glucose intolerant.

Mechanism of action

Metformin decreases hepatic gluconeogenesis and glucose output, improves peripheral glucose uptake and utilization in insulin-sensitive tissue such as muscle and adipose tissue, and may reduce intestinal glucose transport. Its actions depend on the presence of adequate β -Cell function, and therefore insulin in the circulation. The exact mechanism of action of metformin is not clear but it may work through the activation of an intracellular AMP-activated protein kinase.

Lactic acidosis

Lactic acidosis is a rare but serious complication that can occur whilst taking metformin. Lactic acidosis was more frequently reported with phenformin, another biguanide that was subsequently withdrawn in most countries in 1977 after 306 documented cases. While phenformin excretion relies upon hepatic hydroxylation (pharmacogenetically deficient in approximately 10% of Caucasians), metformin is subject to renal tubular secretion, and its excretion depends only on renal function.

Lactic acidosis presents with non-specific symptoms such as lethargy, nausea, vomiting, altered level of consciousness and abdominal pain. Biochemical features of lactic acidosis are those of an elevated anion gap metabolic acidosis with high blood lactate concentrations. There appears to be no direct correlation between blood concentrations of metformin and lactate, and the overwhelming majority of cases of lactic acidosis have occurred in the setting of major comorbidities (Table 16.2). These include conditions associated with increased lactate formation (e.g. cardiac failure, pulmonary disease), reduced lactate metabolism (e.g. hepatic failure) and reduced metformin and lactate excretion (e.g. renal failure).

There is an increased risk of lactic acidosis in patients treated with metformin who receive radiological contrast media. This may be avoided by withholding metformin for a few days before and after the intended radiological procedure.

A systematic review has shown that there is no evidence from prospective comparative trials or from observational cohort studies that metformin is associated with an increased risk of lactic acidosis, or with increased plasma concentrations of lactate, compared with other antihyperglycaemic treatments, if prescribed as licensed. Patients with significant cardiac, renal or liver dysfunction should not receive metformin. As metformin is renally excreted, there is the attendant risk of accumulation in acute or chronic kidney disease, resulting in an increased risk of lactic acidosis. It is recommended by NICE that prescription of metformin should be reviewed in patients with a glomerular filtration rate (GFR) $<45 \text{ mL/min/1.73m}^2$ and it is to be avoided in those with a GFR $<30 \text{ mL/min/1.73m}^2$. Some physicians recommend that metformin should not be used if the patient's plasma creatinine concentration is raised to any extent; others argue that guidelines have been progressively tightened with no reduction in the incidence of lactic acidosis, which has anyway always been very low throughout the history of metformin usage. Although some formularies suggest avoiding metformin when plasma creatinine

TABLE 16.2 Cautions with and contraindications to the use of metformin

Cautions or contraindications	Reason
Old age	Renal and other organ impairment
Tissue hypoxia	Increased lactate production via anaerobic glycolysis, reduced gluconeogenesis
Pulmonary disease	Increased lactate production via tissue hypoxaemia and anaerobic glycolysis
Liver failure, ethanol	Reduced lactate clearance via inhibition of gluconeogenesis
Iodinated radiological contrast media	Risk of acute kidney injury, competition for renal tubular secretion
Cardiac impairment	Increased lactate production via poor tissue oxygen delivery
Shock states, severe dehydration	Increased lactate production via poor tissue oxygen delivery
Advanced microvascular disease	Risk of chronic kidney disease
Severe infection	Increased lactate production
Drugs affecting tubular secretion of metformin	Contrast media, cimetidine, digoxin, others
Type 2 diabetes itself	Microcirculatory disturbances, peripheral vascular disease, increased incidence of cardiac and renal disease, reduced activity of pyruvate dehydrogenase in insulin resistance favouring anaerobic metabolism

concentration is raised above certain thresholds, cardiac dysfunction is probably a more important, but less well recognized, contraindication to its use.

Other unwanted effects of metformin

The major side-effects of metformin include gastrointestinal problems such as nausea, abdominal discomfort, metal taste and diarrhoea. Many patients will experience these symptoms after commencing metformin but will eventually develop tolerance. These side-effects are generally dose dependent, and can be minimized by commencing at a low dose and then slowly titrating the dose upwards, or by using a slow-release preparation. Approximately 80% of the total benefit of metformin is seen with doses of around 1.5 g/day, which is close to the median tolerated dose. Metformin induces malabsorption of vitamin B₁₂ and folate but this is rarely clinically significant.

Sulfonylureas (and related insulin secretagogues)

First generation sulfonylurea (sulphonylurea) derivatives were first used to treat diabetes in the 1950s, following the observation in 1942 that certain sulphonamide antibiotics could provoke severe hypoglycaemia in patients treated for typhoid fever. They act as insulin secretagogues, reducing blood glucose concentrations by augmenting the first-phase release of immediately releasable (reserve) insulin from β -Cells. Unlike glucose-stimulated insulin release, there is little direct effect on insulin production. First generation sulfonylureas, such as chlorpropamide, were beset with unwanted effects including the syndrome of inappropriate antidiuresis (SIADH) and flushing. These drugs have largely been replaced by more potent and shorter acting second and third generation drugs (e.g. gliclazide) that lack these unwanted effects.

Mechanism of action

Central to the release of insulin from secretory granules (but not the only mechanism involved) is the closure of an ATP-sensitive membrane-bound inwardly rectifying 140 kDa K⁺ channel, K_{ATP}. In β -Cells, this channel is a

hetero-octamer of pore-forming units (K_{IR} 6.2) together with regulatory ATP binding cassette (ABC) protein sulfonylurea receptor 1 (SUR-1) molecules. The regulatory subunit, SUR-1, contains 13–17 transmembrane domains with a large number of possible protein kinase A or C phosphorylation sites, together with a number of distinct binding sites for other molecules (including the A and B sites to which sulfonylureas and other exogenous compounds attach, as discussed below). The existence or nature of the endogenous ligand at sulfonylurea binding sites is not clear. Closure of K_{ATP} channels depolarizes the plasma membrane, causing rapid influx of calcium ions via voltage dependent calcium channels. The resultant increase in free ionized cytosolic calcium concentration ([Ca²⁺]_i) triggers cytoskeletal trafficking of secretory granules to the plasma membrane and release of insulin by a process of exocytosis. Amplification of the signal generated by K_{ATP} closure occurs via a number of incompletely understood mechanisms, including direct mobilization of calcium from intracellular stores (e.g. by the action of glucagon-like peptide (GLP)-1 analogues). The effects of K_{ATP} channel opening are terminated by extrusion of K⁺ via a voltage gated K⁺ channel.

The principal mechanism of action of the sulfonylurea drugs is to close the K_{ATP} channel by binding to the A and B sites on SUR-1. Conversely, diazoxide, used in the treatment of certain hyperinsulinaemic conditions, is a potent opener of K_{ATP} channels, and thiazide diuretics also open β -Cell K_{ATP} channels (causing impairment of glucose tolerance).

Adverse effects of sulfonylureas

Concern has existed for many years that sulfonylurea drugs may have harmful cardiovascular effects (mainly arrhythmias). This concern was first raised in relation to tolbutamide (the University Group Diabetes Programme, UGDP). The action of sulfonylureas on cardiac K_{ATP} channels suggests a plausible mechanism for adverse cardiac effects.

In contrast to these predominantly theoretical reservations, analysis of UKPDS data does not support the suggestion of adverse cardiovascular effects from sulfonylureas. At the present time, and in the light of the Diabetes Insulin–Glucose in Acute Myocardial Infarction (DIGAMI) study results, many physicians would agree

that discontinuation of sulfonylurea drugs and substitution with insulin infusion is probably indicated in the setting of acute coronary syndromes. Some would also advocate the use of selective agents (e.g. nateglinide or glimepiride) in other cardiac patients, pending definitive data.

Other unwanted effects of sulfonylureas

Unwanted effects of the sulfonylureas include hypoglycaemia (particularly in elderly patients treated with longer-acting agents and in renal/hepatic failure), serious immunological reactions (Stevens–Johnson syndrome and other rashes), marrow dyscrasias, hepatitis, weight gain (typically 1–3 kg) and precipitation of acute porphyria. Contraindications to the use of these drugs include type 1 diabetes, pregnancy, lactation and significant hepatic and renal insufficiency.

Hypoglycaemia owing to deliberate or inadvertent overdose of sulfonylureas requires prompt and aggressive management. The first-line treatment is with a bolus of 50% glucose followed by continuous infusion of 10% or 20% glucose. If blood glucose cannot be maintained at safe levels with this treatment alone, hydrocortisone, glucagon or octreotide may be useful.

Indications and clinical usage

Sulfonylurea and derivative drugs are used to treat patients with type 2 diabetes who have adequate insulin reserve, but in whom lifestyle and dietetic improvements have failed to control hyperglycaemia. Doses are titrated to blood glucose and HbA_{1c} responses; primary failure of either to improve may indicate advanced β -Cell failure. Because of their tendency to promote weight gain, sulfonylureas are best used as first-line agents in non-obese patients, but may be added to metformin, thiazolidinediones or both in obese patients with secondary failure of these agents. Insulin therapy can be added if and when necessary. Weight gain with repaglinide and nateglinide (see below) may be less than with conventional sulfonylureas, and severe hypoglycaemia may be less common.

The extent to which postprandial hyperglycaemia rather than fasting hyperglycaemia contributes to diabetic complications remains the subject of considerable debate. Nevertheless, because of its ability to augment glucose-stimulated insulin release (as opposed to basal insulin release), meglitinides may be useful when postprandial hyperglycaemia in a patient with type 2 diabetes is the dominant clinical problem.

Meglitinides

Repaglinide and nateglinide are prandial glucose regulators that form the meglitinide class of drug. They are indicated in the treatment of patients with type 2 diabetes who have poor glycaemic control on diet control or metformin therapy. These agents are short-acting and act to increase insulin secretion by closing K_{ATP} channels of the β -Cell membrane via binding to a receptor distinct from that of sulfonylureas. Repaglinide binds to the B site of SUR-1 and nateglinide binds to the A site. The meglitinides are used in patients as monotherapy (repaglinide

only) or in combination with metformin (repaglinide and nateglinide). Given the rapid onset and short duration of action, these agents are taken between 15 and 30 min before each meal to provide postprandial glycaemic control. As repaglinide and nateglinide are rapidly metabolized in the liver and excreted in bile, these agents should not be used in patients with hepatic impairment.

Adverse effects of meglitinides

Hypoglycaemia is a common adverse effect. Other reactions include gastrointestinal effects such as abdominal pain, diarrhoea, constipation, nausea and vomiting; hypersensitivity responses such as pruritus, rashes and urticaria may occur.

Peroxisome proliferator activator γ analogues

Thiazolidinediones were originally developed as ligands for the orphan nuclear peroxisome proliferator activator receptor family (PPAR α , PPAR γ and PPAR δ). The endogenous activators of the nuclear PPAR receptors are fatty acids and fatty acid-derived eicosanoids, and it is now known that the actions of the fibrate group of lipid-lowering agents are mediated via PPAR α receptors. Activation of PPARs leads to the formation of heterodimers with the retinoid X receptor (RXR), bound to its own endogenous ligand, retinoic acid. These PPAR–RXR dimers bind to specific gene regulatory sequences (peroxisome proliferator response elements, PPREs), modulating the transcription of multiple (>40) target genes.

Thiazolidinediones are insulin sensitizers that act as ligands at the PPAR γ receptors, which regulate gene expression and transcription of glucose transporter molecules GLUT-4, lipoprotein lipase, fatty acid transporter protein and fatty acyl CoA synthase. The PPAR γ nuclear receptors are found in adipose tissue and also in liver and muscle. Their activation leads to improved insulin responsiveness by increasing glucose uptake, and adipose lipogenesis by regulating expression of adipose tissue adipokines.

Troglitazone, the first of these compounds to be introduced, was withdrawn following cases of idiosyncratic hepatotoxicity leading to acute liver failure. At present, the only thiazolidinedione in clinical use is pioglitazone. The other drug in this class, rosiglitazone has been linked to an increased risk of cardiovascular disorders such as myocardial infarction and heart failure, and to bone fractures. It is no longer recommended for use.

Mechanisms of action

It has been suggested that the insulin-sensitizing effects of PPAR γ agonists are almost exclusively a result of a ‘fatty acid steal’ mechanism, owing to their effect on adipose tissue, rather than direct effects on the pancreas, liver or peripheral insulin-sensitive tissues. This is because PPAR γ receptors are expressed chiefly in adipose tissue, with only low levels of expression in other insulin-sensitive tissues. Thiazolidinediones, acting via PPAR γ , increase free fatty acid uptake into adipose tissue (by ~60%) and also increase fatty acid oxidation in the liver, heart, kidneys and skeletal

muscle. These effects on adipose tissue are brought about by altering adipocyte gene expression, pre-adipocyte differentiation and fat distribution (favouring redistribution from central to subcutaneous depots). By such actions, hepatic NEFA uptake is reduced (by ~40%), rendering the liver more insulin-sensitive and giving these agents a potential role in the treatment of hepatic steatosis. However, thiazolidinediones have multiple actions, and it is by no means established that 'fatty acid steal' is their dominant mechanism of action.

Peroxisome proliferator activator receptor α is expressed in tissues that metabolize fatty acids extensively such as liver, kidneys, heart, and muscle. Its activation increases plasma HDL-cholesterol concentrations by transcriptional induction of the HDL apolipoproteins, apoA-I and apoA-II (see Chapter 37). Triglyceride concentrations are reduced by decreased hepatic apoC-III production and increased lipoprotein lipase expression. Low density lipoprotein subclass distribution is shifted toward larger particle species by changes in lipoprotein composition and interaction. Combination PPAR α/γ agonists are being developed in the hope that a single tablet may be able to benefit both the adverse lipid and glycaemic profiles of type 2 diabetes.

Pioglitazone may be used alone or in combination with metformin or a sulfonylurea, or with both as 'triple therapy'. While this may delay the need for insulin in some patients, the usual scenario is that pancreatic reserve is failing and that further delays in introducing insulin treatment are potentially hazardous. Pioglitazone is also licensed in the UK for prescription with insulin when metformin cannot be used. The maximum effects of pioglitazone take some 4–12 weeks to develop, and patients may experience a deterioration in glycaemic control during this period if switching from another agent. Pioglitazone is also sometimes used in the treatment of hepatic steatosis (an unlicensed indication). This use, however, should be considered carefully in each patient because of the associated risks.

Adverse effects

As pioglitazone may cause fluid retention, it is contraindicated in patients with heart failure; this risk may be increased if it is used in combination with insulin. The use of pioglitazone is associated with a small increased risk of bladder cancer it is therefore not recommended in patients with a history of bladder cancer or uninvestigated haematuria and should be used with caution in the elderly, given that the risk of bladder cancer rises with age.

Contraindications to the use of pioglitazone also include severe renal insufficiency, pregnancy, breastfeeding and hepatic dysfunction, with the possible exception of that arising from steatohepatitis. Fears over the potential to cause liver damage requires that plasma aminotransferase activity should be monitored two-monthly over the first year after starting treatment. Weight gain of 1–4 kg is predictable from the mechanism of action, but the potential for this to interfere with insulin sensitivity is more than offset by the favourable redistribution of adipose tissue. Fluid retention may cause a mild dilutional anaemia (haemoglobin typically falls by 10–20 g/L) and ankle oedema (in 5–10% of individuals, more when used in combination with insulin).

Glucagon-like peptide 1 analogues

Proglucagon is a product of the glucagon gene, which undergoes post-translational cleavage in pancreatic α -cells to form glucagon and in the L cells of the small bowel, to form glucagon-like peptide 1 (GLP-1), which is produced in response to food in the gastrointestinal tract. Glucose-dependent insulinotropic polypeptide (GIP) is secreted by the K cells in the duodenum. The incretin effect occurs when oral glycaemic stimuli, in the form of food intake, stimulate insulin secretion. It is mediated by the gastrointestinal incretins GLP-1 and GIP, which stimulate insulin synthesis and secretion. In addition, GLP-1 delays gastric emptying and intestinal transit and reduces food intake by promoting satiety.

The incretin response is reduced in type 2 diabetes and the incretin analogues (called incretin mimetics) have been developed for use in its treatment. Human GLP-1 has a short half-life because of the degradation of the N terminal by dipeptidyl peptidase IV (DPP-4) and therefore synthetic GLP-1 analogues (exenatide and liraglutide) have been developed that share the effects of GLP-1 but are resistant to cleavage by DPP-4.

Mechanisms of action

The actions of GLP-1 include increasing the insulin secretion resulting from any given glycaemic stimulus, slowing the rate of gastric emptying and intestinal transit, increasing satiety and reducing the secretion of glucagon. Glucagon-like peptide 1 also has trophic effects on β -Cells, stimulating differentiation and proliferation of progenitor cells in the pancreatic ductal epithelium and inhibiting β -Cell apoptosis. Glucagon and GLP-1 both stimulate insulin secretion by raising cytosolic calcium concentrations by mechanisms distinct from those involving the K_{ATP} channel of β -Cells. Receptor binding increases cyclic AMP generation, which impedes sequestration of calcium into intracellular organelles following glucose-stimulated K_{ATP} channel closure. Endogenous GLP-1 has a very short half-life (1–2 min) and is therefore not a practical therapy.

Exenatide is a synthetic analogue of a 39 amino acid exendin-4 peptide that was found in the saliva of the *Heloderma suspectum* (Gila monster) lizard. It requires twice daily subcutaneous injection. More recently, a once weekly preparation of exenatide has been approved. Liraglutide has a more prolonged duration of action compared to exenatide because of greater resistance to inhibition by DPP-4.

These agents are used in patients with type 2 diabetes who have inadequate glycaemic control on one or two oral agents who are intolerant of other oral agents, or in whom use of other oral agents is contraindicated. Their use should preferably be reserved for patients with a body mass index (BMI) >35 kg/m² or when BMI is <35 kg/m² and insulin therapy is not appropriate for work-related reasons because of the risk of hypoglycaemia, or when weight loss is desirable. Recently both liraglutide and exenatide have also been shown to be of benefit when used in combination with insulin, as they can help reduce insulin doses and promote weight loss.

Adverse effects

The main side-effects that limit the tolerability of GLP-1 analogues are gastrointestinal problems such as nausea, diarrhoea, vomiting and abdominal distension, and injection site reactions. Hypoglycaemia is possible when used in combination with a sulfonylurea or insulin. Renal failure has been reported with exenatide use. Pancreatitis has been reported, and should be considered in patients taking GLP-1 analogues who report severe abdominal pain. Liraglutide should not be used in patients with a history of medullary thyroid carcinoma or multiple endocrine neoplasia type 2 because of potential stimulation of thyroid C-cells, observed in rodent studies.

Dipeptidyl peptidase IV inhibitors

The GLP-1 and GIP incretins are metabolized by cleavage of the N-terminal by dipeptidyl peptidase IV (DPP-4), an enzyme that is ubiquitously expressed. Inhibiting DPP-4 activity prolongs the half-life and increases circulating concentrations of the incretins, thereby stimulating insulin secretion and inhibiting glucagon release. Members of the DPP-4 inhibitor class of drugs include: sitagliptin, saxagliptin, linagliptin and vildagliptin. These may be used as monotherapy or in combination with metformin, a sulfonylurea or pioglitazone in patients who fail to reach glycaemic targets with monotherapy; some can also be used in combination with insulin.

Adverse effects of DPP-4 inhibitors

Inhibitors of DPP-4 do not cause weight gain or weight loss. Hypoglycaemia is uncommon but may occur when they are taken with a sulfonylurea or insulin. Other adverse effects include upper respiratory tract infections, headache and nasopharyngitis. Cardiovascular outcome studies are currently ongoing but thus far, have not shown any significant reduction in cardiovascular risk.

Alpha-glucosidase inhibitors

This class of drugs (acarbose, miglitol) inhibit the α -glucosidase enzymes that break down carbohydrates into monosaccharides at the intestinal border. By delaying the breakdown of carbohydrates in the small bowel, prandial absorption of glucose is reduced, resulting in improved control of postprandial blood glucose concentrations. The use of these drugs is limited by intolerance to the gastrointestinal effects of undigested carbohydrate in the bowel, causing, in particular, excessive flatulence and diarrhoea. The gastrointestinal symptoms do, however, discourage excessive carbohydrate consumption. Hypoglycaemia may occur in association with use of insulin or sulfonylureas.

Sodium-glucose co-transporter 2 (SGLT2) inhibitors

This new class of drugs includes dapagliflozin and canagliflozin. These agents block the reabsorption of glucose by the SGLT2 transporter in the proximal renal tubules, resulting in 70-80 g of glycosuria per day. They have been shown to reduce HbA_{1c} by between 7 and 11 mmol/mol

when used as monotherapy or added to other agents such as metformin and sulfonylureas in dual or triple combinations, or with insulin. They also result in modest weight loss (about 2-3 kg) and reductions in blood pressure. Because of their mode of action, they are less effective in patients with renal impairment.

Side effects include an increased risk of fungal genital infections and a slightly increased risk of urinary tract infections. The intrinsic risk of hypoglycaemia is low. They should be avoided in patients taking loop diuretics because of the small risk of volume depletion.

Insulin use in type 2 diabetes

Insulin may be initiated in patients with type 2 diabetes when glycaemic control is poor or at presentation if symptoms and signs are such that type 1 diabetes cannot be excluded. Patients with type 2 diabetes require insulin treatment after a median of seven years from diagnosis. The decision to start insulin treatment is often a difficult one, and will depend on many factors relating to individual patient preference and circumstances as well as glycaemic status. It is apparent from the UKPDS that type 2 diabetes is a progressive condition, and that once fasting hyperglycaemia occurs, drug treatment, even when intensive, will improve glycaemic control only transiently, with the subsequent deterioration of β -Cell function and glycaemic control paralleling that of less intensively treated patients, albeit after some initial improvement. Many patients are reluctant to start insulin but are willing to take several oral hypoglycaemic agents. In some cases, the addition of a third oral hypoglycaemic (e.g. metformin, DPP-4 inhibitor plus thiazolidinedione or metformin, sulfonylurea plus DPP-4 inhibitor) may be effective, at least for a while. Responses to the addition of a thiazolidinedione are variable, but in some cases worthwhile. However, treatment failures are also common, and failure to reduce HbA_{1c} values by a pre-determined amount (e.g. >10 mmol/mol (1%)) after six months should lead to reconsideration of insulin treatment, as adequate insulin reserve is required for thiazolidinedione action.

Initial insulin treatment in type 2 diabetes is usually with either a once daily intermediate acting insulin (isophane) or long-acting insulin (glargine or detemir). Some patients are commenced on a twice daily premixed insulin regimen (e.g. 30% short-acting insulin and 70% intermediate acting insulin). Insulin treatment in individuals who are overweight may lead to further weight gain. This is because of: the anabolic effects of insulin; reduced energy loss through glycosuria; possibly increased oral intake because of need to avoid or treat hypoglycaemia, and altered lifestyle, exercise and diet with insulin therapy. With increased weight, these patients face problems of increased insulin resistance, and consequently, the required insulin doses needed to maintain glycaemic control rise, leading to further weight gain.

Bariatric surgery

Weight loss is an important component in the management of overweight and obese subjects with type 2 diabetes. Bariatric surgery may be performed in selected obese patients with diabetes, preferably before the development of irreversible complications and after the exhaustion of

conservative measures including diet, exercise and drug therapy. Surgery should be considered in patients with type 2 diabetes who have a BMI of $>35 \text{ kg/m}^2$ (as recommended by NICE and the ADA). It is likely that this will be an increasingly considered option for patients with suitable psychological and medical reserves (see Chapter 11 for further details).

Bariatric surgery improves glycaemic control, depending on the surgical procedure, with 'remission' rates of diabetes (defined as resolution of clinical and laboratory manifestations of type 2 diabetes) reported to be as high as 78% in a meta-analysis. Surgery is associated with restoration of first-phase insulin response in up to 80% of subjects. The Swedish Obesity Study compared the effects of bariatric surgery in 2010 obese subjects with 2037 obese controls and found that bariatric surgery was associated with a decreased incidence of diabetes and, in those with type 2 diabetes at baseline, increased diabetes remission rates at two years and at ten years. Not only have improvements in glycaemia been reported well before weight loss becomes apparent, but they appear also to exceed those expected on the basis of the amount of weight lost. It has been proposed that increased secretion of a number of gut peptides with insulinotropic actions such as GLP-1 and GIP, and decreased secretion of orexigenic peptides such as ghrelin, may be responsible for the success of bypass techniques, raising the intriguing possibility that operations of this kind have a partly endocrine mechanism of action. As energy intake following surgery is changed, it may be necessary to adjust glucose lowering therapy to avoid severe hypoglycaemia.

Pancreatic transplantation

Simultaneous pancreatic and kidney transplantation has been performed in patients with established renal failure, and is a reasonable option to consider at this stage in patients with glycaemic instability, who are physically fit and psychologically well prepared. Pancreas-alone transplants have also been carried out and may sometimes be considered in patients with recurrent hypoglycaemia. Although complete independence from insulin medications is possible, this form of treatment requires immunosuppressive therapy and is limited by the availability of organs for donation.

Islet cell transplantation

Islet cell transplantation may be used in patients with severe hypoglycaemia unawareness. This uses cadaveric pancreatic islets cells that are infused through a catheter placed into the hepatic portal vein. However, it is limited by donor availability and low rates of independence from insulin medications because of graft attrition. The Clinical Islet Transplantation (CIT) consortium, a network of centres in North America and Europe, is conducting studies on islet transplantation in type 1 diabetes, but the clinical outcomes are not yet known.

Immunotherapy for type 1 diabetes

Immunomodulatory therapies for type 1 diabetes are the subject of phase II and phase III clinical trials. These

treatments target specific antigens or modulate the immune system with the aim of establishing immune tolerance and preventing the destruction of pancreatic β -Cells. The concept of tolerance is important in that in autoimmune disease, the immune system loses its tolerance to a particular antigen, which may be a mechanism underlying the autoimmune nature of type 1 diabetes. Different immune-targeted treatments that target pathogenetic cells and activate regulatory cells have been developed. Nutritional studies have investigated how exposure to different dietary components may influence disease initiation.

Glutamate decarboxylase (GAD65) immunization of young non-obese diabetic (NOD) mice has prevented development of type 1 diabetes. However, in humans, there was no effect on β -Cell destruction. Inoculation with human heat shock protein 60 (Hsp60) prevented disease progression in NOD mice, whilst in humans, it produced increased C-peptide concentrations. Studies in children have not shown an improvement in β -Cell function, however. Trial evidence for anti-CD3 and anti-CD20 monoclonal antibodies therapy in humans has shown improved C-peptide concentrations and reduced insulin demand. Monthly injections with CTLA4-Ig (a T cell receptor that inhibits T cell activation) over two years in human trials also increased C-peptide concentrations. Trials of interleukin-2 showed favourable results in NOD mice. Trials with insulin therapy have not shown effects on cell function.

Nutritional studies in children showed that lower concentrations of islet cell antibodies were present in those who had been weaned in infancy onto a hydrolysed casein formula compared with those weaned to cows' milk formula. A trial of the effects of ω -3 fatty acids showed reduced risk of islet autoimmunity in children with a genetic risk of type 1 diabetes. In another trial, vitamin D did not protect β -Cell function in patients with recent onset of type 1 diabetes.

OBSTACLES TO ACHIEVING GLYCAEMIC CONTROL

Glycaemic management of diabetes is rarely as simple as making incremental additions to a patient's therapy until normoglycaemia is achieved. Hypoglycaemia is the main barrier to tight glycaemic control. For the majority of patients, achieving good glycaemic control involves a complex trade-off between the potential benefits (most of which are long term) and risks or obstacles (many of which are only too readily and immediately apparent).

The patterns of endogenous insulin secretion, which vary rapidly in response to changes in blood glucose concentration, are not easily mimicked by the injection of exogenous insulin into a subcutaneous depot, from which there is prolonged release into the systemic rather than portal circulation, thus partially bypassing hepatic action and first-pass clearance. This leads to an inherently increased risk of hypoglycaemia, which may be exacerbated by a number of factors related to the treatment, the individual and his or her circumstances.

Intensive control

The ADA recommends an HbA_{1c} concentration of <53 mmol/mol (<7%) as a treatment goal whereas the International Diabetes Federation (IDF) recommends an HbA_{1c} of <48 mmol/mol (<6.5%).

The DCCT demonstrated that in type 1 diabetes, intensive insulin treatment (multiple daily injections or CSII therapy) improved glycaemic control and reduced retinopathy, nephropathy and neuropathy complication rates. Subsequently, the Epidemiology of Diabetes Interventions and Complications (EDIC) study showed that this effect persisted for at least ten years in those subjects who had been in the intensive glycaemic control group of the DCCT.

In type 2 diabetes, the UKPDS showed reduced microvascular complications with intensive glucose control. At long-term follow-up at ten years, early intensive glycaemic control continued to reduce the risk of microvascular complications. However, a recent meta-analysis demonstrated that intensive glycaemic control in type 2 diabetes is associated with a 30% increase in the relative risk of severe hypoglycaemia. Several large trials have compared the effects of intensive versus standard glycaemic control on cardiovascular outcomes in type 2 diabetes. The Action in Diabetes and Vascular Disease: Preterax and Diamicon Modified Release Controlled Evaluation (ADVANCE) trial showed a significant reduction in microvascular outcomes (nephropathy and retinopathy) with tightly controlled diabetes but no significant reduction in cardiovascular outcomes. The Veterans Affairs Diabetes Trial (VADT) trial found that intensive glycaemic control did not significantly lower the incidence of cardiovascular events compared with standard glycaemic control. The Action to Control Cardiovascular risk in Diabetes (ACCORD) trial found increased mortality associated with an intensive glycaemic control to a target HbA_{1c} of <42 mmol/mol (<6%), leading to the early termination of the trial.

In both type 1 and type 2 diabetes, the effect of intensive control of blood glucose on microvascular complications is therefore well established. Although the ACCORD, ADVANCE and VADT trials did not

show any improvement in cardiovascular outcomes with intensive control, glycaemic goals should still be set because of the adverse microvascular consequences associated with poor glycaemic control. Both the ADA and the International Diabetes Federation recommend individualized targets for each patient and the need to balance the risks and benefits of tight control of HbA_{1c}. Goals for patients with a history of frequent or severe hypoglycaemia, long standing diabetes, the elderly or those with limited life expectancy and patients with multiple comorbidities will need to be carefully considered. The established cardiovascular risk reduction measures such as blood pressure control, smoking cessation, lipid lowering therapy and healthy lifestyle should continue to be recommended as part of an individualized care plan.

Hypoglycaemia

The risk of hypoglycaemia often limits the achievement of good glycaemic control in diabetes. Although microvascular and macrovascular complications of diabetes typically evolve over a period of several years, hypoglycaemia may occur quickly, over a few minutes, and the consequences may be devastating or permanent. Some 2–4% of deaths and a much greater proportion of quality adjusted life years (QALYs) lost in type 1 diabetes are attributable to hypoglycaemia. Furthermore, the perceived danger, inconvenience, embarrassment and helplessness associated with hypoglycaemia is considerable, and the extent to which an individual patient will risk hypoglycaemia often influences how they manage their glycaemic goals and how they view care for their diabetes. For some patients, the risk of hypoglycaemia conflicts with occupational use of a motor vehicle (and hence possibly their livelihoods). In other patients (e.g. those with cancer or the very elderly), the primacy of goals such as freedom from osmotic symptoms and hyperglycaemic emergencies do not necessitate tight glycaemic control.

In addition to the factors mentioned in Table 16.3, it is clear that reduced glucoregulatory responses to hypoglycaemia in patients with type 1 diabetes and longstanding

TABLE 16.3 Factors that may affect the effect of insulin on blood glucose concentration

Treatment-related factors	Comments
Insulin	Dose, chemical form (human, animal, analogue, lipophilia, tendency to form complexes, albumin binding), suspension (protamine zinc etc.), site and volume of injection (lipohypertrophy, exercising limbs vs abdominal wall), timing in relation to meal (30 min before, immediately before, after), time of day (insulin resistance higher in the morning), changes in insulin clearance (decreased GFR), accidental intravenous injection of insulin, warming (e.g. shower soon after injection), weather or massage of the injection site
Diet and absorption	Glycaemic index, carbohydrate and fat content, missed meals, gastric absorption (presence of gastroparesis, drugs causing rapid transit), changes in absorption (coeliac disease), extended interval between meals (e.g. overnight, special circumstances) and dialysis (glucose removal)
Gluconeogenesis and glycogenolysis	Depletion of hepatic glycogen reserves after fasting, inhibitory effects of alcohol or other drugs, liver disease
Insulin resistance	Time of day, associated conditions and their treatment (Addison disease, hypothyroidism), exercise, intercurrent illness, sleep disturbance, mental/emotional/social stress, weight loss, improved metabolic control (glucotoxicity), obstetric delivery or intrauterine fetal death (sudden reduction in insulin resistance), cessation of steroid therapy

type 2 diabetes can markedly increase the frequency and severity of hypoglycaemic episodes.

In persons without diabetes, the first response to a falling blood glucose concentration is a reduction in insulin secretion; this typically occurs at blood glucose concentrations <4.5 mmol/L. This response is not available to subjects with type 1 diabetes or to many subjects with type 2 diabetes (either insulin or secretagogue treated). Glucagon secretion forms the next layer of defence, stimulating hepatic glycolysis and gluconeogenesis. However, most patients with either type 1 or type 2 diabetes are chronically hyperglucagonaemic and cannot respond to hypoglycaemia in this way.

The last level of defence against acute hypoglycaemia is activation of the sympathetico–adrenomedullary system, which normally occurs when blood glucose falls to <2.8 – 3.0 mmol/L. This increases lipolysis and circulating NEFA production and utilization, and mobilization of substrates for gluconeogenesis further inhibits insulin secretion and acts as a stimulus to glucagon release. Importantly, activation of the sympathetico–adrenomedullary system gives rise to the first clear symptoms of hypoglycaemia. With still lower blood glucose concentrations, patients also experience symptoms of neuroglycopenia. Although neuroglycopenia usually manifests as a generalized phenomenon, focal neurological deficits may occur in the setting of relatively normal consciousness and, for this reason, capillary blood glucose testing must be performed (and any low result confirmed by formal blood glucose measurement) in all cases of apparent stroke, confusion or other acute neurological disability, and in patients presenting with unexplained seizures.

Patients with diabetes are advised to carry dextrose (glucose) tablets or quick acting carbohydrate in case of emergency. Parenteral correction of hypoglycaemia may be necessary in cases where the patient has lost consciousness or is not able to eat food. An intramuscular glucagon injection may be used to correct insulin-induced hypoglycaemia. In severe cases of hypoglycaemia, for example where the patient is unconscious, intravenous glucose may need to be administered by medical personnel. Patients who have had recurrent episodes of hypoglycaemia should be reviewed by the diabetes team to assess potential causes and ways to avoid this from happening.

Hypoglycaemia-associated autonomic failure

Inability to reduce insulin secretion and failure of the glucagon response are described above. At this point, only the sympathetico–adrenomedullary responses stand between the patient and the development of neuroglycopenia. Failure of sympathetico–adrenomedullary responses can also occur and will lead both to a reduced awareness of, and reduced conscious and unconscious response to, hypoglycaemia (hypoglycaemia unawareness). Its particular importance is that it is associated with a greatly increased risk (possibly by a factor of 25) of severe hypoglycaemia. This potentially serious situation is usually amenable to treatment (see below).

Drugs such as β -blockers may interfere both with the action of adrenaline (epinephrine) and also with the perception of sympathetico–adrenomedullary activation,

and should be used with great caution in the context of hypoglycaemia unawareness. Autonomic neuropathy may lead to reduced sympathetico–adrenomedullary function and reduced hypoglycaemia awareness. It is also associated with unpredictable glycaemic excursions owing to erratic intestinal transit times. Perhaps most importantly, functional impairment of sympathetico–adrenomedullary responses may occur in type 1 or long-standing type 2 diabetes in response to preceding hypoglycaemia. Recurrent hypoglycaemic episodes lead to altered body responses to hypoglycaemia and susceptibility to further episodes.

Hypoglycaemia-associated autonomic failure (HAAF) may occur after a single episode of hypoglycaemia, although this is more frequently seen following recurrent hypoglycaemia. There is evidence to support a role for cortisol in mediating HAAF. The hypothalamo–pituitary–adrenal axis is stimulated by initial bouts of hypoglycaemia, but the resulting hypercortisolaemia may result in impaired adrenaline (epinephrine), noradrenaline (norepinephrine) and glucagon responses to subsequent hypoglycaemia. Muscle sympathetic nerve activity, an index of sympathetic nervous system outflow, is also impaired following hypoglycaemia. Infusion of pharmacological doses of cortisol in healthy subjects has a similar effect on their subsequent hormonal response to insulin-induced hypoglycaemia. However, this effect has not been replicated using lower dose infusions that achieve plasma cortisol concentrations similar to those found in hypoglycaemia, and hypoadrenal subjects without diabetes who are treated with fixed physiological cortisol replacement doses demonstrate almost complete preservation of counter-regulatory responses to a subsequent episode of hypoglycaemia, calling this hypothesis into question. Regardless of the underlying mechanism, however, management of HAAF is based upon avoidance of hypoglycaemia for several days, by slightly relaxing glycaemic control, thereby allowing autonomic responses to recover.

The principle of antecedent glycaemia affecting subsequent responses is further illustrated by the converse phenomenon of ‘clinical pseudohypoglycaemia’. This term denotes the observation that some patients with chronic hyperglycaemia experience symptoms of hypoglycaemia at blood glucose concentrations well in excess of those that would trigger responses in individuals with lower prevailing blood glucose concentrations. Management of this condition is empirical, and should involve measures to induce very gradual improvements in glycaemic control.

The Somogyi effect and the dawn phenomenon

After a counter-regulatory response to a hypoglycaemic episode (whether or not followed by the administration of (sometimes excessive) glucose to reverse the symptoms), rebound hyperglycaemia is common. Usual glycaemic control, glucose tolerance and insulin sensitivity may take several hours to be restored. A special variant of the rebound from hypoglycaemia is the Somogyi phenomenon, which occurs in patients with nocturnal hypoglycaemia. The patient may not be aware of the hypoglycaemia, even in retrospect, although awakening with malaise, headache and bedclothes damp from sweating are suggestive. The

rebound from the nocturnal hypoglycaemia results in the patient waking with a blood glucose concentration higher than desirable, provoking a temptation to take at least as much (or even more) insulin the next night and thereby increasing the likelihood of a subsequent episode of hypoglycaemia. Detection of this problem requires a high level of suspicion and may involve measures such as setting the alarm clock in order to self-test, the observations of another person, or continuous blood glucose monitoring.

Subjects without diabetes show circadian changes in blood glucose concentration and glucose tolerance based upon changes in counter-regulatory hormones, insulin secretion and the availability of alternative fuels for metabolism. The most marked such circadian effect is the 'dawn phenomenon', which typically occurs between 04.00 h and 07.00 h and is an increase in plasma glucose concentration and decrease in insulin sensitivity consequent upon the increased secretion of counter-regulatory hormones at that time. During this period, people with diabetes usually experience modest rises (1–2 mmol/L) in plasma glucose concentrations without recent ingestion of food. A larger rise is seen in patients in whom circulating insulin concentrations are too low at this time.

Owing to both the dawn phenomenon and the Somogyi effect, there is a risk of hypoglycaemia earlier in the night whenever longer-acting insulins, particularly when administered in the early evening, are titrated to pre-breakfast blood glucose concentrations.

Exercise

A patient's usual balance between carbohydrate intake and insulin requirements may be influenced by physical activities. The metabolic effects of exercise vary with the intensity and duration of that exercise. For example, jogging typically elicits an increase in both glucose and lipid oxidation, and plasma glucose and NEFA turnover increase. In more strenuous exercise, muscle stores of glycogen and triglycerides will become depleted. Plasma glucose concentrations may increase or decline during exercise, depending on the counter-regulatory hormonal drive to hyperglycaemia, and whether glycogen reserves are adequate.

In healthy subjects, plasma insulin concentrations decrease, permitting increased hepatic glucose output to limit the fall in blood glucose. In subjects on exogenous insulin treatment, insulin concentrations may not fall (indeed, there may be a rise if increased skin blood flow during exercise increases insulin absorption from subcutaneous injection sites) and hence hepatic glucose output may not be sufficient to prevent hypoglycaemia. After exercise, there is a phase of glycogen repletion in liver and muscle, and in subjects with diabetes, delayed hypoglycaemia may be a problem at any time up to several hours afterwards.

Strategies to control blood glucose concentrations during exercise include adjusting insulin doses that are to be taken prior to planned exercise, or ingestion of extra carbohydrate before, during or after the exercise. Because of the numerous physiological variables involved, it is

often difficult to predict the best strategy in advance, and a system of trial and error with increased frequency of blood glucose monitoring may be necessary. Many patients with diabetes lead active lifestyles as regular physical exercise is important in improving insulin sensitivity and improving and maintaining cardiorespiratory fitness.

Ethanol

Ethanol (alcohol) may inhibit gluconeogenesis; in patients with diabetes, this may reduce hepatic glucose output sufficiently to provoke hypoglycaemia. Typically, this does not occur while the patient is consuming ethanol, since most beverages contain adequate carbohydrate to prevent this (or the patient eats as well as drinks): instead, hypoglycaemia occurs during the first two hours after alcohol consumption. However, the inhibition of gluconeogenesis persists for several hours and a rebound hypoglycaemic episode may occur a few hours (typically 2–4) after consuming alcohol. If the clinical features of hypoglycaemia are dismissed as the consequences of inebriation, then the failure to treat the hypoglycaemia may be serious. In addition, the effects of alcohol may impair the judgement of an individual with diabetes, who may forget to consume carbohydrate after heavy alcohol consumption or fail to adjust their insulin dose in anticipation of planned alcohol consumption. It may even be a contributory cause of ketoacidosis when insulin doses are unintentionally omitted.

Intercurrent illness, 'sick day rules' and stress

Intercurrent illnesses such as viral infections usually cause hyperglycaemia in people with diabetes but may also give rise to hypoglycaemia. The effect on glycaemic control may persist for up to two weeks, and insulin requirements, even during simple infections (e.g. colds), may increase by 100%, although increases of ~20% are perhaps more typical.

It is likely that many episodes of hyperglycaemic crisis could be prevented by appropriate self-treatment and timely medical advice during times of intercurrent illness. Treatment plans should ideally be individualized and prepared in advance. Several general principles apply to all, however. These include the concepts that:

- glycaemic treatment generally requires augmentation rather than reduction during times of intercurrent illness
- frequent (e.g. four-hourly) blood glucose testing is needed
- medical advice should be sought if blood glucose concentrations become and remain high despite appropriate measures (e.g. >13–15 mmol/L)
- maintenance of good hydration is of the utmost importance.

For patients with type 2 diabetes, metformin should probably be discontinued for the duration of any severe illness, and a sulfonylurea (or, if necessary, insulin) substituted in the short term. Patients taking a sulfonylurea may require dose increments or the short-term use of insulin.

Patients with type 1 diabetes are at risk of ketoacidosis and should therefore be advised to test their urine regularly (at least twice daily) to ensure no more than light to moderate ketonuria is present. Where blood ketone testing is available, testing for blood ketones is preferred. Patients should be warned against the common misconception that if they are not eating (e.g. because of vomiting), they should stop taking insulin. Instead, essential fluid and glucose may be ingested in the form of regular small volumes of soup, rehydration fluid or proprietary sugar-containing drinks. Usual doses of basal insulin are continued and the day's bolus (rapid-acting) insulin divided into 4-hourly quotas, which can be adjusted on the basis of self-testing. Persistently elevated or rising blood glucose concentrations, particularly in the context of increasing ketonuria, should prompt contact with an appropriate medical advisor or admission to hospital for insulin infusion.

In patients with severe intercurrent illness such as septicæmia, pancreatitis or myocardial infarction (all these are more common in diabetes), intravenous insulin is often required, even in subjects not usually requiring exogenous insulin treatment. There are now good data to support the use of intravenous insulin infusion in such circumstances and during major surgery. It is not yet clear whether the infusion of insulin per se is of benefit or whether tight glycaemic control achieved by any means confers similar advantage. These issues are discussed in greater detail in a later section.

Emotional stress of many forms will result in changes in secretion of counter-regulatory hormones and autonomic tone, which will alter glucose tolerance. Since emotional stress is often associated with alterations in sleep, eating and exercise patterns, it may be practically impossible for patients with diabetes under emotional stress to maintain usual glycaemic control.

CHRONIC COMPLICATIONS OF DIABETES

The complications of diabetes can be divided into acute and chronic complications. Acute complications include hypoglycaemia (discussed above), DKA and the hyperosmolar hyperglycaemic state (see p. 326). Chronic complications include macrovascular disease (see above), retinopathy, nephropathy and neuropathy.

Nephropathy

Diabetic nephropathy is the leading cause of established renal failure and the most common diagnosis at initiation of renal replacement therapy. Glomerular damage is pathologically described as either diffuse or nodular glomerulosclerosis (Kimmelstiel–Wilson lesions). The pathological process involves thickening of the basement membrane and of the glomerular mesangium with the accumulation of extracellular collagen matrix material. There is glomerular podocyte loss and tubulointerstitial fibrosis. Glomerular hyperfiltration is followed by microalbuminuria, proteinuria and subsequent decline in renal glomerular function.

Some 20–30% of patients with type 1 diabetes will develop renal disease (most commonly 15–25 years after diagnosis). Risk factors for its development include longer duration of disease, an earlier age at diagnosis, earlier onset of puberty, poorer glycaemic control throughout the time course of the disease, smoking and a family history of diabetic nephropathy. Nephropathy is probably less prevalent among those with type 2 diabetes (10–20% lifetime risk), but seems likely to increase substantially with the increasingly early mean age of onset of type 2 diabetes. Diabetic nephropathy is screened for by laboratory testing for urine albumin excretion (urine albumin/creatinine ratio), serum creatinine and estimated glomerular filtration rate (eGFR). Screening should be performed at least annually. (Formulae used for the calculation of eGFR are discussed further in Chapter 7.)

Microalbuminuria

The early stages of the disease are asymptomatic and include hyperfiltration (with urine albumin excretion (UAE) <30 mg/24 h or 20 µg/min) followed by progression through microalbuminuria (UAE 30–300 mg/24 h or 20–200 µg/min) to proteinuria (UAE >300 mg/24 h or 200 µg/min). After proteinuria has become established, the glomerular filtration rate (GFR) begins to fall and progression to established renal failure is generally inevitable, although the rate of loss of renal function may still be usefully slowed by treatment.

Differences in historical ten-year progression rates from microalbuminuria to nephropathy (~80%), and more recent estimates (~30%), lend weight to the hypothesis that aggressive treatment at the stage of microalbuminuria may be able to halt or even reverse the disease in many patients. Thus, the first and best opportunity to detect the disease clinically is at the stage of microalbuminuria. Some would even contend that risk associated with UAE is a continuous variable, with detectable increases in morbidity above the normal median value of albumin excretion of approximately 2.5 mg/24 h. Conventional dip-stick testing of urine is not usually positive at such concentrations of albumin excretion. Neither this test nor the measurement of 24 h urinary protein excretion are now recommended as screening tools. Screening should probably take place at the time of diagnosis and annually thereafter in type 2 diabetes. Although it has traditionally been thought that screening could safely be delayed for five years in type 1 diabetes, a number of studies have shown significant prevalence before five years, especially where puberty has intervened or in the presence of poor glycaemic control, and many physicians now advocate annual screening from diagnosis in this group also.

Factors giving rise to falsely elevated results include unaccustomed exercise, intercurrent illness, other renal disease, haematuria, acute deterioration in glycaemic control and urinary tract infection (which should be excluded whenever tests of microalbuminuria are positive, see Chapter 8). Furthermore, considerable day-to-day variations in urine albumin excretion rates mean that a minimum of two spot urine collections for

albumin:creatinine ratio over a period of 3–6 months should be abnormal before a diagnosis of microalbuminuria can be made reliably. It should be noted that the occurrence of non-immunoreactive forms of albumin in the urine may result in underestimation of the true extent of microalbuminuria in some assays. Intermittently positive results, especially where occurring frequently without a discernible cause, do appear to increase the risk of progression to definite microalbuminuria. It is important to appreciate that microalbuminuria is not just a risk factor for the development of nephropathy, but is also an independent risk factor for coronary artery disease (indeed, one of the most potent risk factors known), being also associated with dyslipidaemia, hypertension, endothelial dysfunction and diabetic retinopathy. Patients with microalbuminuria should, therefore, undergo intensive risk factor modification, and low-dose aspirin is generally indicated for these patients. There is controversy as to whether an increase in blood pressure precedes or is a result of the development of microalbuminuria.

Management

Prevention of diabetic kidney disease is of paramount importance. It necessitates good metabolic control, adequate treatment of hypertension, avoidance of smoking, lipid-lowering therapy and avoidance of nephrotoxic drugs. Intensive glycaemic control delays onset and progression of albuminuria. However, even with these measures, data from the UKPDS and DCCT show that a significant number of patients will still develop the disease, albeit at a later stage and with slower progression. Thus, the second important feature of management remains early detection of raised UAE, which is discussed in detail above.

Good metabolic control may delay the onset of microalbuminuria and, once this is present, may halt progression or even reverse it. In general, antihypertensive therapy, maintaining systolic BP <140 mmHg and diastolic <80 mmHg, irrespective of the agent used, slows the development of diabetic nephropathy. The UKPDS provided evidence that blood pressure control could protect against nephropathy. Inhibitors of the renin-angiotensin system, i.e. angiotensin-converting-enzyme (ACE) inhibitors or angiotensin receptor antagonists (ARBs) confer superior long-term protection, probably because of a concomitant specific reduction of transglomerular filtration pressure, preventing increases in albumin excretion and reducing the severity of proteinuria. Angiotensin-converting-enzyme inhibitors reduce cardiovascular events and ARBs reduce progression of nephropathy. Combination ACE inhibition and ARB therapy lowers albuminuria further, but may increase the risk of hyperkalaemia and evidence for the clinical efficacy of this approach is limited. Increases in plasma creatinine concentrations of up to 20–30% from baseline are common on starting treatment and, although they do not mandate alternative treatments, monitoring of renal function is essential.

Dietary protein restriction to 0.8–1.0 g/kg body weight per day in early chronic kidney disease (CKD) or 0.8 g/kg body weight per day in advanced CKD is recommended.

Blood pressure control with other agents such as diuretics, calcium channel blockers and β -blockers may be used.

End-stage disease

Despite prevention, detection and aggressive treatment of diabetic nephropathy, a significant proportion of patients will nevertheless progress towards established renal failure (ERF). Early referral to a nephrology service (usually before GFR has fallen to <30 mL/min/1.73 m²) will usually be appropriate. Occult cardiac disease is common in this group of patients, but its angiographic investigation is hazardous owing to the risk of contrast nephropathy. Many patients with anaemia other than that caused by specific haematinic deficiencies will benefit from erythropoietin treatment at haemoglobin concentrations of <110 g/L. Management of metabolic sequelae such as secondary hyperparathyroidism, acidosis and hyperphosphataemia are, in general, as for patients with non-diabetic kidney disease. Treatment options for ERF include haemodialysis and renal transplantation, depending on the availability of graft donors and compatibility.

Neuropathy

Neuropathy in diabetes has many different manifestations and can be focal or diffuse. The types of neuropathy include chronic sensorimotor, autonomic, diabetic amyotrophy (proximal motor neuropathy) and mononeuropathies. An acute neuropathy may also occur in the presence of hyperglycaemia (hyperglycaemic neuropathy), possibly as a result of damage to peripheral nerves.

Chronic sensorimotor neuropathy

This is most the common type of neuropathy in patients who have had diabetes for a long time. The longest nerves are usually affected, with sensory loss in a glove and stocking distribution. The sensory loss is often accompanied by loss of power and autonomic involvement. There is loss of pain and light touch, with decreased reflexes and muscle wasting in affected parts. Some patients may report painful symptoms (e.g. tingling, burning, hyperalgesia or allodynia).

Patients should be screened for diabetic neuropathy by pinprick tests, using a 10 g monofilament and a 128 Hz vibration fork applied to the distal aspects of both feet. The diagnosis is often made clinically without the need for electrophysiological studies, although these can be performed if the diagnosis is uncertain. It is important to screen for other, potentially reversible, causes of neuropathy, e.g. vitamin B₁₂ deficiency, alcohol excess, heavy metal poisoning and neurotoxic medications, and for conditions that may require specific treatment, e.g. demyelination. Patients with sensorimotor neuropathy are at increased risk of injury to and ulceration of the feet. Charcot neuroarthropathy (Charcot foot, see below) is a common complication. All patients with peripheral neuropathy should receive regular podiatric assessment and advice regarding foot care because of the increased risk of foot ulceration.

Besides optimizing blood glucose control, the treatment for sensorimotor neuropathy includes symptomatic pain relief with agents for neuropathic pain such as duloxetine, amitriptyline, pregabalin and gabapentin.

Autonomic neuropathy

Autonomic neuropathy is seen in patients with diabetes who have longstanding poor control. Manifestations include orthostatic hypotension (a major risk factor for falls), constipation, gastroparesis (with nausea and vomiting), loss of normal Valsalva response, gustatory sweating, resting tachycardia, bladder dysfunction, erectile dysfunction and altered autonomic responses to hypoglycaemia.

Diabetic autonomic neuropathy may affect gastrointestinal motility: this condition may be occult, or manifest as constipation, diarrhoea, unexplained (or easily triggered) vomiting or malabsorption. Altered gastrointestinal motility may change the rate of absorption of nutrients from food. Rapid and unpredictable absorption is particularly unwelcome in diabetes and may completely undermine the attempts of even a well-motivated patient to maintain good glycaemic control. In some cases, diabetic ketoacidosis may be provoked as a result of repeated vomiting. If gastrointestinal motility can be improved, this renders glycaemic control somewhat easier; however, such amelioration is usually spontaneous rather than related to medical intervention. Investigations for gastroparesis include the use of isotope scintigraphy to assess solid phase gastric emptying. Prokinetic treatments such as metoclopramide, domperidone or erythromycin may be helpful in some cases to aid motility. Dietary changes may be necessary and in severe situations, parenteral nutrition may become necessary if adequate nutrition is not possible by the enteral route. A gastric pacemaker with electrodes that electrically stimulate gastric contractions may be used in severe cases.

Patients who are troubled with postural hypotension may be prescribed support stockings. In severe cases, fludrocortisone (a mineralocorticoid) or midodrine (an α -adrenergic agonist) may be used.

Autonomic neuropathy may impair hypoglycaemia awareness. This will often result in the patient (very sensibly) avoiding the possibility of severe hypoglycaemia and, as a result, glycaemic control may be more difficult to achieve. Autonomic neuropathy may be associated with increased morbidity and mortality if it affects the cardiovascular system. This is manifested by postural hypotension and resting tachycardia and is associated with an increased risk of cardiovascular events.

Erectile dysfunction and retrograde ejaculation can occur in autonomic neuropathy. If the sacral nerves innervating the bladder are affected, patients may suffer from bladder dysfunction, with bladder emptying difficulties, incontinence problems and urinary tract infections. Erectile dysfunction may be caused by multiple factors such as vascular disease, neuropathy, medication or psychological problems. Besides reviewing these factors, treatments for erectile dysfunction may include phosphodiesterase type 5 inhibitors, prostaglandins and vacuum devices (see Chapter 23).

Mononeuropathies

Mononeuropathies may result from diabetes affecting single nerves or, where multiple nerves are affected, a mononeuritis multiplex. These include third cranial nerve palsies (that may have a vascular aetiology), lateral popliteal nerve palsy and carpal tunnel syndrome.

A form of multiple mononeuropathy called diabetic amyotrophy may occur. This is caused by a reduction in the blood supply to the nerves of the lower limbs, for example the femoral nerve or the lumbar plexus, resulting in proximal weakness and wasting of muscles in the lower limbs. Amyotrophy is associated with poor glycaemic control, and optimum glycaemic control is important in attaining recovery, which is a slow process. Pain control and physiotherapy are helpful.

The feet in diabetes

Foot ulcers

Foot ulceration in patients with diabetes is a common cause of morbidity and its complications, such as sepsis, may be associated with mortality. Ulcers mainly occur in patients who have both diabetic neuropathy and peripheral vascular disease. Repeated trauma without recognition by the patient may cause infection. Foot architecture may become distorted because of callus formation, deformity and ulceration, leading to a condition called the Charcot foot (see below).

Concomitant infection of the feet does not heal well owing to repeated unrecognized foot trauma and poor vascular circulation. Infective organisms include *Staphylococcus aureus*, *Streptococcus pyogenes* and *Pseudomonas aeruginosa*.

Charcot foot

Charcot foot is a specific foot deformity, bilateral in approximately 20% of cases, which may arise on the background of severe sensory (and probably motor) neuropathy. Often precipitated by minor trauma, the disease may progress very rapidly over weeks or a few months, beginning with an unstable, warm, red and swollen foot (stage 0), the point at which treatment should be initiated for maximum benefit. A 'fragmentation' stage (stage 1) follows, with periarticular fractures, joint dislocation, instability and deformity. In the 'coalescence' stage (stage 2), reabsorption of bone debris occurs before finally a stable (stage 3), but deformed, foot is left at the end of the process.

The exact pathogenesis of this condition remains unclear, but abnormal protective motor reflexes may permit the development of abnormal loading forces within the foot which, in the presence of sensory neuropathy and osteopenia, may damage bone, joints and ligaments. Subluxation, or even dislocation, of the metatarsal, tarso-metatarsal or tarsal joints may occur and swelling may be caused by leakage of synovial fluid from joint capsules. Bone blood flow is usually substantially increased, perhaps as a result of abnormal sympathetic control, and this may contribute to the osteopenia. Later in the course of the disease, bones may become sclerotic. Fractures may

cause the tarsal bones to collapse, resulting in an outward bowing of the arch ('rocker-bottom foot' deformity), further adding to the abnormal loading and risk of ulceration. The foot can be so warm and red that osteomyelitis, cellulitis or gout is suspected. Osteomyelitis may be particularly difficult to distinguish when a plantar ulcer is present and, indeed, may be very likely if there are systemic signs of infection or if bone is encountered when the ulcer is probed. Plain radiographs usually show vascular calcification and bone changes; the latter may be difficult to distinguish from osteomyelitis. Magnetic resonance imaging or bone biopsy may be required to establish the correct diagnosis in some cases.

Intravenous bisphosphonates (e.g. pamidronate), total immobilization and offloading of the foot (e.g. bed rest, pending the fitting of a total contact cast, perhaps for four months or longer) and meticulous attention to metabolic control are helpful in the acute setting. Bone markers, such as serum carboxy-terminal telopeptide of type 1 collagen, a marker of osteoclastic bone resorption, measured in venous blood taken from the dorsum of the foot, urinary deoxypyridinoline and bone-specific alkaline phosphatase, have been used in studies alongside longitudinal measurements of foot temperature to assess the response to treatment.

Eye disease

Diabetes is a common cause of visual loss, especially in the developed nations. Risk factors for the development of diabetic retinopathy include duration of diabetes, poor glycaemic control (chronic hyperglycaemia), diabetic nephropathy, hypertension, hyperlipidaemia, smoking, obesity and rapidly improving glycaemic control, for example in pregnancy or when commencing insulin pump therapy. Retinopathy is usually present after 20 years of diabetes, although the severity varies between patients. Patients with diabetic retinopathy are predisposed to cataracts, glaucoma and macular disorders.

Patients with type 1 diabetes should have an eye examination within five years of diagnosis. The prevalence of retinopathy in type 1 diabetes after 10–15 years is 25–50%. All patients with type 2 diabetes should be screened when diagnosed, reflecting the fact that the presence of type 2 diabetes may have not been recognized for some time. About 60% of patients with type 2 diabetes have non-proliferative retinopathy. All patients should be screened annually, or more frequently depending on the findings. Women with diabetes are screened more frequently during pregnancy and in the postpartum period because of an increased risk of progression of diabetic retinopathy.

Eye examinations should be carried out by an ophthalmologist or optometrist experienced at diagnosing diabetic retinopathy. In the UK, an annual retinopathy screening service is provided for all patients with diabetes. Any patient with progressive retinopathy detected at screening is referred to an ophthalmologist for follow-up investigation and, if necessary, treatment. The UK National Retinal Screening Committee classification is based on the presence or absence of both retinopathy (graded from R0 to R3) and maculopathy (graded M0 or M1) (see [Box 16.1](#)).

BOX 16.1 The UK national screening programme for diabetic retinopathy classification system

- R0 – No retinopathy
- R1 – Background changes (microaneurysms, exudates not involving the macula)
- R2 – Preproliferative changes (venous abnormalities, intraretinal microvascular anomalies (IRMA), soft exudates)
- R3 – Proliferative retinopathy (new vessel formation, vitreous haemorrhage, rubeosis iridis)
- M0 – No referable maculopathy
- M1 – Maculopathy requiring referral to ophthalmology (macular haemorrhages, macular exudates and macular oedema or ischaemia)

Initially, there is an increase in retinal perfusion. Background retinopathy denotes the formation of microaneurysms and hard exudates. Flame-shaped or dot and blot haemorrhages may follow should the microaneurysms rupture. Retinal capillaries may leak. Leakage into the macular region may cause maculopathy. Maculopathy is characterized by the formation of hard exudates in close proximity to the fovea, along with oedema and ischaemia of the macula region. It is more common in type 2 than type 1 diabetes and is the commonest cause of loss of vision.

Preproliferative retinopathy is diagnosed on the basis of the presence of cotton wool spots (soft exudates), intraretinal microvascular anomalies and venous loops with arterial occlusion. Soft exudates occur when there is infarction of the nerve fibre layer. Intraretinal microvascular anomalies result from dilated capillary beds from capillary occlusion. Continued retinal ischaemia and hypoxia resulting from small vessel occlusion triggers the release of angiogenic factors such as vascular endothelial growth factor (VEGF). Neovascularization of the retinal vasculature leads to proliferative retinopathy. These newly formed vessels are vulnerable to rupture, leading to intraocular vitreal haemorrhages and visual loss. The neovascularization is associated with fibrous tissue that may cause traction on, and detachment of, the retina or cause haemorrhage from vessel tear. New vessel formation on the iris is known as rubeosis iridis and may lead to glaucoma. Proliferative retinopathy is more common in type 1 diabetes with a prevalence of ~25% after 15 years.

Tight glycaemic and blood pressure control reduces the incidence and progression of diabetic retinopathy. However, improving glycaemic control should be a gradual process as a transient worsening of retinopathy has been reported with rapid improvements in glycaemic control. Smoking should be discouraged. Treatment with xenon or argon laser photocoagulation is effective in reducing visual loss in proliferative retinopathy and maculopathy. The photocoagulation aims to destroy hypoxic retinal tissue. Pan-retinal photocoagulation may be performed in patients with severe proliferative retinopathy. Focal laser photocoagulation is applied to macular oedema and leaking vessels. However, this treatment cannot reverse vision that has already been lost prior to treatment. Following therapy, some patients experience

visual field constriction and night blindness. Vitrectomy may be carried out in patients with vitreous haemorrhage and this improves visual recovery. New therapies for macular oedema currently under investigation include intravitreal steroid injections and use of VEGF antagonists. Intravitreal steroid injections may be used where there is persistent loss of vision and when conventional treatment has failed. This therapy is effective but the adverse effects include an increased risk of cataracts and raised intra-ocular pressure. Concerns about chronic VEGF inhibition relate to neurotoxicity and thromboembolic events, such as myocardial infarction, stroke and thromboembolism.

Other complications

Brittle diabetes

Brittle diabetes is a term with no universally accepted definition. In 1977, Tattersall used a clinical definition, describing 'The patient whose life is constantly being disrupted by episodes of hypo- or hyperglycaemia, whatever their cause'. This definition has been modified by various workers.

Causes of brittle diabetes include: psychological abnormalities, such as eating disorders (e.g. bulimia), personality disorders, communication disorders and manipulative behaviour. Such factors are often inferred if good glycaemic control can be achieved in hospital by the use of intravenous or nurse-administered subcutaneous insulin. If intravenous insulin therapy fails to achieve good glycaemic control, then an organic cause should be sought. These include: an unsuitable insulin regimen, intercurrent illness such as thyroid disease, Addison disease, systemic lupus erythematosus (through antibodies to insulin or its receptor), disorders of intestinal motility (erratic absorption of food) and interactions with other drugs (prescribed or non-prescribed). Problems related to insulin injections, such as faulty technique, abnormalities of local anatomy (e.g. lipohypertrophy) or subcutaneous tissue blood flow (e.g. during exercise) and changing insulin requirements in the various phases of the menstrual cycle, are well-recognized causes.

Type 4 renal tubular acidosis

Hyporeninaemic hypoaldosteronism (type 4 renal tubular acidosis, RTA) may be a manifestation of diabetic nephropathy, particularly in older patients. It presents with a hyperchloraemic, hyperkalaemic metabolic acidosis, which is out of proportion to any impairment in renal function. Type 4 RTA is frequently unmasked by inhibition of the renin-angiotensin system by ACE inhibitors or ARBs, particularly when used in combination with other agents that may increase plasma potassium concentrations (e.g. spironolactone). The precise nature of the defect in this condition is unclear. Failure of renin secretion to increase in response to posture or sodium restriction suggests an interstitial (juxtaglomerular) defect, but no anatomical correlate has been described to account for this. The failure of aldosterone release to be stimulated directly by the resulting hyperkalaemia suggests the possibility of concomitant dysfunction of the adrenal zona

glomerulosa, although this may simply be a consequence of chronic hyporeninaemia.

When hyperkalaemia is modest and stable (serum potassium concentration <6.0 mmol/L), close observation may be all that is required, but more marked elevations demand measures such as stopping offending drugs (ACE inhibitors/ARBs, β -blockers, non-steroidal anti-inflammatory drugs, potassium-sparing diuretics etc.); prescription of a low potassium diet (potassium restriction); use of potassium-wasting diuretics (thiazides, loop diuretics) and/or sodium bicarbonate (although this risks causing oedema). A differential diagnosis of autoimmune hypoadrenalism may require exclusion (especially in younger subjects with type 1 diabetes). In the occasional particularly severe and refractory case, treatment with fludrocortisone may be necessary.

EMERGENCIES IN DIABETES

General medical emergencies (e.g. stroke) in persons with diabetes are usually treated much in the same way as in subjects without diabetes, although frequent monitoring of blood glucose is necessary. Specific diabetic emergencies are discussed as specific entities, although mixed forms may occur. Diabetic emergencies are common in clinical practice and the approach to such patients requires an accurate diagnosis of the condition and appropriate management.

Diabetic ketoacidosis

Diabetic ketoacidosis (DKA) is a serious complication of type 1 diabetes and can be life-threatening. It is characterized by hyperglycaemia, acidosis and ketonaemia. Approximately 30% of patients with type 1 diabetes mellitus present with ketoacidosis. Fortunately, the onset in newly presenting patients tends to be slower than in known diabetic subjects, perhaps because of retention of some residual insulin secretion. An infection is found in ~35–55% of patients with ketoacidosis. Errors or omissions in treatment account for another 30%. The clinical features include dehydration, shock, vomiting, abdominal pain, acidosis (with compensatory Kussmaul respiration) and, in some cases, a breath odour of ketosis) and impaired consciousness.

Four mechanisms predispose to ketoacidosis: insulin deficiency, counter-regulatory hormone excess, fasting and dehydration. Of these, the most important is insulin deficiency. Mechanisms of the development of DKA are shown in [Figure 16.1](#). In a vicious cycle, hyperglycaemia and excess lipolysis cause dehydration and high circulating concentrations of NEFAs. The resulting acidosis and dehydration further increase the secretion of counter-regulatory hormones. The acidosis, dehydration, counter-regulatory hormones and excess lipolysis induce insulin resistance, thereby increasing hyperglycaemia and lipolysis. Protein catabolism and hypertriglyceridaemia also occur and compound the situation.

Rarely, patients may present with normoglycaemic ketoacidosis; typically, this occurs in situations where

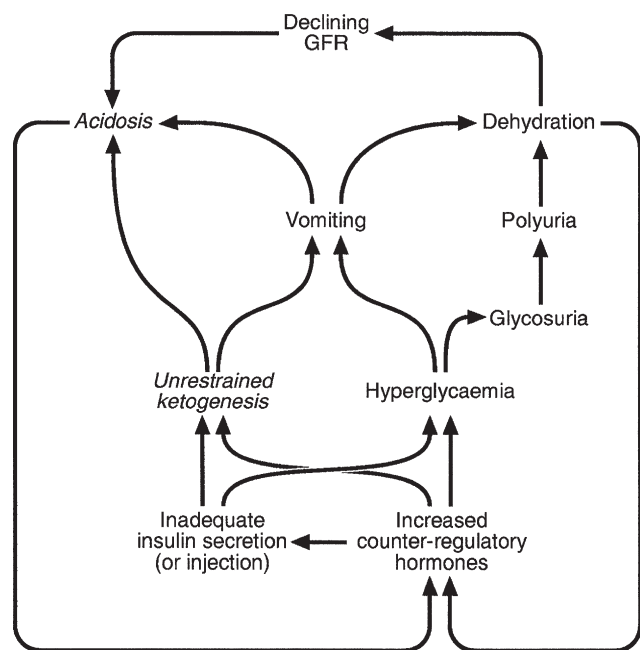


FIGURE 16.1 ■ Vicious circles in ketoacidosis and hyperosmolar states. In both conditions, the trigger event(s) causes an initial inadequate insulin supply and hyperglycaemia. In ketoacidosis, both acidosis and dehydration cycles occur. Features in italics are restricted to ketoacidosis. GFR, glomerular filtration rate.

exercise, starvation or, occasionally, infection is the major precipitant of ketosis, but more often it is seen when a patient has taken enough insulin in a bolus to bring the blood glucose down but has not been exposed to sufficient insulin over a long enough period to suppress ketogenesis while existing ketones are cleared. The important differential diagnosis of alcoholic ketoacidosis should be considered under these circumstances (see below), particularly when a malnourished patient embarks on heavy alcohol consumption for a period of several days. Despite the unfortunate use of the term ‘diabetic coma’, it is rare for a patient with DKA to have true coma as opposed to some clouding of the sensorium. Where this occurs, alternative causes such as bacterial meningitis or a cerebrovascular event should be actively considered in the differential diagnosis.

Biochemical features

Biochemical features of DKA include hyperglycaemia, ketosis (ketonaemia and ketonuria), metabolic acidosis and uraemia. Creatinine concentrations may not accurately be measurable in the presence of heavy ketonaemia. The characteristic ketosis is a consequence of increased lipolysis and decreased fat synthesis. Excess acetyl-CoA derived from β -oxidation of fatty acids is converted to the ketone bodies, acetoacetate and β -hydroxybutyrate with some decarboxylation of the former to acetone.

Bedside tests may demonstrate high blood glucose concentrations, glycosuria and ketonuria. Plasma β -hydroxybutyrate concentrations are typically three times higher than those of acetoacetate and the ratio may be even greater if there is also an element of lactic acidosis, so that the degree of ketosis may be underestimated

or not recognized by urine tests for acetoacetate (see Chapter 15). This is also important in charting the recovery from ketoacidosis where ketonuria (detection of acetoacetate) may increase at a time when the patient is, by all other measures, improving and ketonaemia resolving. This effect occurs as mitochondrial redox state improves under conditions of increased tissue oxygen delivery and reduced acidosis, favouring the formation of acetoacetate from β -hydroxybutyrate. Some typical biochemical results in ketoacidosis are shown in [Box 16.2](#).

Management

The key to successful management of ketoacidosis is early institution of therapy and repeated clinical and biochemical assessment of the patient. The essential elements of the treatment of ketoacidosis are to replace fluid losses and to reverse the underlying metabolic disorder with continuous intravenous infusion of insulin. Ketone clearance occurs mainly through renal excretion and oxidation.

General measures. As with other medical emergencies, initial stabilization with attention to airway, breathing and circulation should precede detailed assessment. It is recommended that venous (rather than arterial) blood gases and hydrogen ion concentration (pH), and bicarbonate and potassium concentrations should be used to guide management. Appropriate investigations including blood culture and chest X-ray should be undertaken where infection is suspected, but the finding of a neutrophil leucocytosis does not, of itself, imply its presence. Although abdominal pain and a slightly elevated amylase activity

BOX 16.2

Typical initial laboratory values in diabetic ketoacidosis

- Plasma [glucose] ~35 mmol/L
- Plasma [K^+] >5.3 mmol/L: whole body depletion typically 6.0 mmol/kg body weight
- Plasma [Na^+] ~130 mmol/L: whole body depletion typically 8.0 mmol/kg body weight
- Plasma [urea] >15 mmol/L
- Plasma [creatinine] >150 μ mol/L (if possible to measure in the presence of heavy ketonaemia)
- Plasma [ketones] (β -hydroxybutyrate and acetoacetate) >15 mmol/L
- Plasma [Mg^{2+}] <0.70 mmol/L: whole body depletion typically 0.5 mmol/kg body weight
- Plasma [phosphate] >1.2 mmol/L: whole body depletion typically 1.0 mmol/kg body weight
- Plasma amylase 500–1000 U/L
- Plasma osmolality ~325 mmol/kg: whole body water depletion typically 75–100 mL/kg body weight, i.e. 7 L in typical adult
- Arterial blood gases:
 - [H^+] >50 nmol/L (pH <7.30)
 - $PaCO_2$ <3.52 kPa
 - PaO_2 >12 kPa
 - [HCO_3^-] <18 mmol/L
- Anion gap ($[Na^+] + [K^+] - ([Cl^-] + [HCO_3^-])$) >20 mmol/L

do not necessarily point to the presence of pancreatitis or other intra-abdominal emergency, the possibility that DKA may have been precipitated by abdominal disease should be borne in mind.

Fluids. Fluid replacement aims to restore circulatory volume (and thereby increase the rate of ketone clearance) and correct electrolyte imbalances. There is disagreement as to the optimal rate and type of fluid replacement in DKA, partly based on fears of provoking cerebral oedema. In the UK, 0.9% saline is commonly used as the fluid of choice for initial replacement as this is available readily and comes with ready mixed potassium at the required concentrations if this is required. The disadvantage of using 0.9% saline is that of hyperchloraemic metabolic acidosis, which may lead to a slower recovery from acidosis because of renal vasoconstriction. Hartmann's solution is often less readily available and contains a fixed, relatively low, potassium concentration.

Paediatric guidelines recommend cautious fluid replacement over 48 h. For adults, initial fluid replacement is usually rapid in the first few hours, but this should be done with caution in young adults (see below, where their greater risk of cerebral oedema is explained).

When blood glucose concentrations fall below 14 mmol/L, it is important to commence an infusion of 10% glucose to prevent hypoglycaemia while the insulin infusion is continued (normoglycaemia is often attained before the ketosis and acidosis resolves).

United States' authorities advocate a 'correction' of measured serum sodium concentration for the dilution that results from osmotic movement of intracellular water to the hyperglycaemic interstitial and intravascular compartments. Various formulae have been suggested, but perhaps the most widely accepted is to add 1.6 mmol/L to measured serum sodium concentration for every 5 mmol/L that the plasma glucose concentration exceeds 5 mmol/L. If the corrected sodium starts to rise above the reference range, a switch to 0.45% saline should be considered. The possibility that the hypertriglyceridaemia sometimes encountered in DKA may give rise to pseudohyponatraemia should also be considered, particularly if the plasma is visibly lipaemic.

In general, fluid replacement of ~6 L over the first 24 h with attention to correcting shock very rapidly at the outset, and thereafter reducing the replacement rate, is appropriate for most patients. Polyuria, if still present, will not abate immediately after the start of therapy, so it is important to infuse enough fluid to achieve a positive fluid balance.

Insulin. The objective is to provide sufficient insulin for suppression of ketogenesis without causing precipitate falls in blood glucose or unnecessary hypokalaemia as a result of the insulin-driven intracellular movement of potassium. In practice, the insulin infusion rate required to achieve this objective will vary between patients, and over time in the same patient, as ketone clearance reduces ketone body-induced insulin resistance. However, a starting rate of 0.1 U/h per kg body weight is typical. The rate can be increased if recovery is slow. Initial bolus insulin doses are no longer recommended.

Patients who are normally on a long-acting basal insulin (detemir or glargine) should continue on their subcutaneous injections in order to provide a basal supply of insulin; this will facilitate better glycaemic control when they eventually revert back to their normal subcutaneous insulin regimens when the intravenous insulin infusion is discontinued.

Previously used variable insulin infusion rates ('sliding scales') are no longer used. They were designed to correct blood glucose as rapidly as possible, rather than to permit sufficient insulin to be infused to suppress ketogenesis for as long as it takes to oxidize or excrete the ketones already present in the circulation. The danger was, therefore, that large insulin doses were given in the first few hours, resulting in either too rapid a fall in blood glucose, and the need to switch to 10% glucose before sufficient saline replacement had been given or in a need to reduce the infusion rate to a level that is insufficient to suppress ketogenesis.

It should be noted that the major determinants of blood glucose reduction early in DKA are expansion of its volume of distribution and restoration of a normal glomerular filtration rate, rather than insulin-mediated glucose uptake into insulin-sensitive cells.

Potassium, magnesium and phosphate. Although whole body potassium depletion is universal in DKA, the initial serum potassium concentration may be low, normal or high. Potassium should not be given until it is clear that the patient is passing urine and renal function has been assessed. Potassium concentrations may fall precipitously with insulin treatment and frequent monitoring is necessary, especially during the early stages of therapy. In general, initial replacement rates of 5–20 mmol/h (typically 10 mmol/h), depending on the serum potassium concentration, are appropriate if the serum potassium is within the reference range. Replacement should be withheld if the potassium concentration is >5.5 mmol/L; if it is <3.3 mmol/L, a higher rate of infusion may be required (with cardiac monitoring), and insulin infusion may need to be suspended.

Phosphate and magnesium concentrations may also fall after the start of treatment. Although there is no hard evidence to support the routine use of phosphate replacement in DKA, there is good reason to suppose that severe hypophosphataemia (i.e. <0.35 mmol/L) may be deleterious. Consequences of severe hypophosphataemia include respiratory and other muscle weakness, haemolytic anaemia and reduced cardiac systolic function; these probably result from reduced generation of ATP because of substrate limitation. Severe hypophosphataemia is more likely to occur when DKA is superimposed upon chronic malnutrition. It seems reasonable to suggest that severe, symptomatic, hypophosphataemia should be treated under such circumstances, e.g. with 20 mmol in 1 L intravenous fluid over several hours. Some authorities recommend that potassium replacement be given as a mixture of the chloride (2/3) and phosphate (1/3) salts throughout.

Bicarbonate. The use of bicarbonate to correct acidosis is not recommended. This is because of the risks of hypokalaemia, hypocalcaemia, paradoxical increase in

cerebrospinal fluid acidosis, worsening intracellular acidosis and hypoxia (via a left shift in the oxygen dissociation curve), a possible increase in cerebral oedema risk (see below), delay in the fall in blood lactate concentration and increased ketogenesis.

Cerebral oedema. Cerebral oedema is one of the most feared complications of ketoacidosis, occurring particularly frequently in children and adolescents in whom it is the leading cause of death in the condition. It occurs in ~1% of cases of DKA in children and adolescents and carries a mortality of up to 90%, despite aggressive treatment with mannitol, intubation and mechanical ventilation. The most consistent associations with the subsequent development of cerebral oedema are more severe acidosis at presentation, the larger amounts of insulin administered in the first hour and higher volumes of fluid given in the first four hours. Contrary to popular suspicion, no association with initial serum sodium concentration (whether or not corrected for glucose) has consistently been shown, although it has been suggested that those who develop cerebral oedema may have a tendency to lower corrected sodium concentrations as treatment progresses.

Resolution. Once serum bicarbonate exceeds 18–20 mmol/L, the focus of management should revert to control of blood glucose, and intravenous insulin is continued using a sliding scale. Subcutaneous insulin is recommenced once the patient is eating and does not require intravenous insulin for other reasons (e.g. sepsis). As discussed above, an apparent increase in ketosis, on the basis of urine dip-stick testing for ketones, is not necessarily significant, and need not direct therapy at this stage if all other parameters are satisfactory.

After an episode of ketoacidosis, patients should always be considered insulin dependent and treated with insulin (even if they were not originally). Steps to reduce the risk of further episodes should be taken if appropriate (e.g. reinforcement of education, change of therapy).

Hyperosmolar hyperglycaemic state

Presentation and clinical features

The hyperosmolar hyperglycaemic state (HHS) was formerly known as hyperosmolar non-ketotic hyperglycaemia. The clinical features are dehydration, severe hyperglycaemia (blood glucose usually >30 mmol/L) and a hyperosmolar state (serum osmolality >330 mmol/kg). By definition, ketosis is not significant with urine ketones of less than '+' or blood ketones <1.0 mmol/L. Blood hydrogen ion concentration is usually <50 nmol/L (pH >7.3) with a normal bicarbonate, although coexistent lactic acidosis from sepsis or myocardial infarction may lead to more severe acidosis. Patients may present with osmotic symptoms such as polyuria, polydipsia, thirst, weight loss and vomiting, and are dehydrated. Drowsiness and coma may occur as a result of the high plasma osmolality; inadequate volume replacement leads to acute kidney injury, and mortality is higher than in DKA.

The HHS occurs mainly in older subjects with type 2 diabetes and most cases occur above the age of 60 years.

About half of patients have not been previously diagnosed with diabetes. Many cases arise from late presentation of diabetes with coincident cardiovascular events or infections (most commonly pneumonia or urinary tract sepsis). The cycle of hyperglycaemia, dehydration and increased counter-regulatory hormones, coincident with a history of osmotic symptoms, is similar to that which occurs in DKA, but may be more severe because of later presentation owing to the absence of significant ketoacidosis. The characteristic hypernatraemia (see [Box 16.3](#)) is caused by renal sodium resorption in response to hypovolaemia, together with osmotic diuresis and glycosuria causing persistent free water loss. The HHS usually occurs in subjects with only marginal insulin deficiency, and their insulinaemia allows sufficient glucose uptake and anti-lipolytic effect to prevent the lipolytic and ketotic problems seen in ketoacidosis.

Management

The management of HHS is similar to that of DKA. Fluid replacement is the mainstay of treatment but careful monitoring is necessary in order to avoid fluid overload, which may result in complications such as pulmonary oedema. Insulin requirements are usually less in HHS than in DKA. The plasma glucose concentration typically declines relatively quickly after the institution of insulin therapy, since patients are usually less acutely insulin-resistant than ketoacidotic patients. Rehydration will expand the volume of distribution of glucose, lowering its concentration, and restoration of a good urine output will permit glucose to be excreted in the urine in significant quantities (up to some 250 mmol/L). Together with insulin-independent uptake into tissues, the result is that blood glucose concentration tends to fall fairly rapidly with initial resuscitation, even if no insulin is administered. Rapid falls in plasma glucose or osmolality may predispose to cerebral oedema by encouraging the movement of water from plasma into the central nervous system, and most authorities recommend that the blood

BOX 16.3 Typical initial laboratory values in the hyperosmolar hyperglycaemic state

- Glucose ~60 mmol/L
- Plasma $[K^+]$ ~4.0 mmol/L: whole body depletion typically 10.0 mmol/kg body weight
- Plasma $[Na^+]$ ~155 mmol/L: whole body depletion typically 8.0 mmol/kg body weight
- Plasma [urea] ~55 mmol/L
- Plasma [creatinine] ~400 μ mol/L
- Whole body water depletion typically 80–120 mL/kg body weight, i.e. 8 L in typical adult
- Arterial blood gases:
 - $[H^+]$ <50 nmol/L (pH >7.30)
 - $PaCO_2$: normal
 - PaO_2 : normal (low if venous thromboembolism or pneumonia)
 - $[HCO_3^-]$ ~18 mmol/L (typically mildly elevated blood lactate)
- Anion gap <20 mmol/L

glucose concentration should if possible be reduced at a rate of no more than ~ 3 mmol/L/h. For this reason, insulin infusion is only required at very low rates (e.g. 0.5–1 U/h) in the initial stages of treatment. Controversy exists as to the ideal choice of fluid replacement. While some advocate the use of hypotonic fluids (e.g. 0.45% saline) after restoration of adequate intravascular volume, others argue that normal (0.9%) saline is, in effect, hypotonic to the patient's plasma and is less likely than hypotonic saline to cause rapid falls in plasma osmolality that may increase the risk of cerebral oedema.

One of the major problems in the treatment of HHS is the severe hypernatraemia that often supervenes in the hours after the initiation of therapy. Commencing treatment with fluid and insulin usually induces a decrease in plasma glucose and urea concentrations, and, therefore, osmolality. However, hypernatraemia typically develops during this period, since the falling osmolality allows water to re-equilibrate into dehydrated cells while free water loss through the kidneys continues. Since there is less acidosis and more profound dehydration, plasma potassium deficits are typically less than those occurring in DKA. The intravascular volume depletion increases plasma viscosity, resulting in a hypercoagulable state and stroke, venous thrombosis and myocardial infarction are common causes of mortality. Heparin is, therefore, recommended as prophylaxis against thrombosis. Comorbidities that may have precipitated the episode should be treated; for example, infections may require antibiotic treatment.

After successful treatment of HHS, the severity of residual glucose intolerance may be very mild and the patient can often be successfully managed on diet alone, although many will be discharged on insulin initially. Patients should be educated on steps to avoid provoking factors in the future. At all stages, it is important to involve the diabetes team in the management of patients with HHS.

Other metabolic acidoses

In diabetic subjects, the vast majority of episodes of acidosis are diabetic ketoacidosis. Other metabolic acidoses are comprehensively discussed in Chapter 5. The differential diagnosis of a metabolic acidosis in a diabetic subject may occasionally be troublesome if one forgets that patients with diabetes may sometimes develop acidosis as a result of other conditions, such as severe infection, renal failure and salicylate poisoning. Diagnostic problems may also arise where there is a mixed picture (e.g. diabetic ketoacidosis and lactic acidosis secondary to sepsis, or alcoholic ketoacidosis occurring in a patient with diabetes). Metformin-associated lactic acidosis and type 4 renal tubular acidosis have been discussed earlier in this chapter.

Alcoholic ketoacidosis

Alcoholic ketoacidosis is included here, as it occasionally presents a difficult differential diagnosis from DKA (particularly given that the latter may occasionally present with an only marginally elevated blood glucose concentration) and the treatment is different in several important

regards. This condition presents with ketoacidosis and low, normal or slightly elevated plasma glucose concentrations (rarely >12 mmol/L). Patients usually have longstanding poor nutritional intake with an alcohol binge of several days' duration, often associated with vomiting.

Ketosis is promoted by lack of insulin action in the setting of starvation, permitting mobilization of NEFAs that provide the substrate for ketone body formation as an alternative fuel. Low insulin concentrations prevent ketone body utilization by insulin-sensitive tissues and dehydration impairs ketone excretion by kidneys. Hypercortisolaemia and high growth hormone and catecholamine concentrations enhance the fatty acid release and hepatic ketogenesis consequent upon insulin deficiency.

Ketogenesis is potentiated by inhibition of hepatic metabolism of acetylcoenzyme A, owing to the action of counter-regulatory hormones such as glucagon, cortisol and catecholamines secreted in response both to the hypoglycaemia and also to extracellular fluid volume contraction (which also restricts urinary ketone elimination). In addition, alcohol metabolism depletes cellular nicotinamide adenine dinucleotide (NAD⁺). An increased ratio of NADH to NAD⁺ increases the proportion of β -hydroxybutyrate relative to acetoacetate. It also restricts pyruvate formation from lactate, leading to accumulation of lactate and depletion of pyruvate, a gluconeogenic substrate. As is the case in DKA, alterations in mitochondrial redox state favour β -hydroxybutyrate over acetoacetate production. A complex acid–base disorder ensues from the combined effects of ketosis causing metabolic acidosis, and a combination of extracellular fluid contraction and vomiting causing metabolic alkalosis, with the final [H⁺] (pH) thus not necessarily reflecting the severity of the metabolic derangements present. Treatment differs from diabetic ketoacidosis in that insulin is not required (except if patients have coexistent diabetes), and glucose infusion (which *must* be preceded by thiamine repletion) forms the mainstay of therapy, together with fluid and electrolyte replenishment.

MANAGEMENT OF DIABETES IN THE HOSPITAL SETTING

The management of diabetes in the hospital setting presents a number of challenges and has the potential to affect patient outcomes. Good glycaemic control has an established role in promoting recovery from infection, improving wound healing, and, in obstetric delivery, to prevent neonatal hypoglycaemia. In the first Diabetes Mellitus Insulin–Glucose Infusion in Acute Myocardial Infarction (DIGAMI 1) Study, insulin-based compared with conventional glucose-lowering management improved survival in patients with acute myocardial infarction and type 2 diabetes. The benefit of insulin treatment was not confirmed in the second DIGAMI trial (DIGAMI 2), in which insulin was associated with an increased risk of nonfatal cardiac events, while metformin appeared to be protective against risk of death. Therefore, the type of glucose-lowering treatment used may influence the clinical outcome.

There is mixed evidence concerning mortality and the use of intensive insulin therapy in order to achieve tight glycaemic control in the hospital setting. Recent data suggest that the use of intensive insulin therapy does not reduce mortality or length of hospital stay compared with less strict control, and may increase the risk of hypoglycaemia. This may reflect the increased prevalence of 'stress hyperglycaemia' in hospitalized patients, or be because altered insulin action *per se* in the seriously ill affects the prognosis.

Many patients admitted to hospital will require changes to be made to their diabetes treatment, at least temporarily. These changes may be made necessary by the need to fast, the use of agents such as glucocorticoids or inotropes (which can directly affect blood glucose concentration), when receiving radiographic contrast material (requiring care in the use of metformin), where there is temporarily increased risk of acute kidney injury or other conditions associated with lactic acidosis (also with metformin, see Table 16.1) or during myocardial ischaemia (possibly posing a risk with some sulfonylureas). Some patients whose blood glucose concentrations are elevated during an acute illness, but who either were not previously known to have diabetes or whose blood glucose concentrations subsequently return to levels below the threshold for diabetes diagnosis may have 'stress hyperglycaemia'.

Hospitalization presents the opportunity for a review of overall diabetes management and education: indeed, for some patients, such an opportunistic intervention may represent their only formal contact with diabetes services. It also presents an opportunity to screen for diabetes in a relatively high-risk population.

The perioperative management of patients with diabetes is to minimize changes in glycaemia brought about by the stress of surgery. Most hospitals will have formal protocols for glucose control pre- and post-surgery. Patients with type 1 diabetes will need to be on insulin at all stages to prevent diabetes ketoacidosis. Patients with type 2 diabetes will usually be able to manage on their oral agents, but may require insulin depending on the nature of the surgical procedure. Following surgery, once oral intake is adequate, most patients will be able to recommence their original treatment regimens.

PREGNANCY

For both mother and fetus, diabetes can profoundly affect the pregnancy in terms of metabolism and outcomes. Specific factors influencing the management of women who are pregnant include the presence of coexisting illness, whether they have gestational or pre-existing diabetes and whether they are likely to deliver in the relatively near future.

With pre-existing diabetes, appropriate preconceptual counselling and care is vital if a baby is planned so that glycaemic control and blood pressure are optimized and medications not recommended in pregnancy (for example statins, ACE inhibitors) are replaced. The ADA recommends a preconception HbA_{1c} of <53 mmol/mol (7%). Lipid control is by dietary measures. Folic acid is prescribed to reduce the risk of neural tube defects. These measures increase the chances of successful conception

BOX 16.4 Potential complications for the infant associated with diabetes in pregnancy

- Congenital malformation
- Polyhydramnios
- Macrosomia
- Intrauterine growth restriction
- Intrauterine mortality and still birth
- Neonatal hypoglycaemia
- Postpartum mortality
- Potentially, diseases such as obesity and diabetes in later life

BOX 16.5 Potential complications for the mother associated with diabetes in pregnancy

- Progression of diabetic retinopathy
- Worsening of gastroparesis
- Progression of diabetic nephropathy
- Complications at delivery owing to infant macrosomia

and lower the risk of complications such as congenital anomalies and miscarriage (Boxes 16.4 and 16.5).

In pregnancy, regular surveillance is important: glycaemic control should be monitored with regular measurement of blood glucose concentrations. Pregnancy induces a state of insulin resistance, leading to hyperglycaemia. This can lead to fetal malformation, macrosomia with attendant delivery complications, increased perinatal mortality and premature delivery. Strict glycaemic targets are set which increases the risk of hypoglycaemia. Retinal screening is important in patients because of the increased risk of progression of diabetic retinopathy. Tight preprandial and postprandial glycaemic targets are advised: NICE recommendations are <5.9 mmol/L fasting and <7.8 mmol/L after meals.

Insulin requirements generally rise during pregnancy because, as the pregnancy progresses, there is a reduction in insulin sensitivity, driven by human placental lactogen. This change serves to divert nutrients and the products of maternal lipolysis (fatty acids and ketones) to the fetus. It is important to recognize that a sudden fall in usual insulin requirements may be a sign of changes in placental function that herald impending delivery, and close liaison with the obstetric team is required.

During labour, it is important to monitor the mother's blood glucose concentrations; maternal hyperglycaemia is treated by intravenous insulin and dextrose infusions. On delivery of the placenta, insulin resistance dramatically declines towards non-pregnant values. Patients with type 1 diabetes and those with type 2 diabetes who were previously on insulin pre-pregnancy revert to their pre-pregnancy insulin regimens. Those with pre-existing type 2 diabetes previously on oral glucose-lowering treatment who wish to breast feed may wish to continue on either metformin or insulin (if they were taking this during pregnancy). Patients with gestational diabetes or diet-treated pre-existing type 2 diabetes do not require any medication postpartum, unless other factors, for example

sepsis or surgery, dictate otherwise. Metformin is safe whilst breastfeeding. Arrangements should be made for women with gestational diabetes to have an oral glucose tolerance test approximately six weeks postpartum if pre-discharge blood glucose values are satisfactory. Women with gestational diabetes have an approximately 50% lifetime risk of developing type 2 diabetes, and should be advised on diabetes prevention and recognition as well as being screened regularly for life.

CONCLUSION

Successful clinical management of diabetes mellitus requires an understanding of the effects of diet and exercise on glycaemic control. Type 2 diabetes is responsible for most cases of diabetes, and as ~80% of patients will die from macrovascular disease, it is essential to provide optimal management of all cardiovascular risk factors, including dyslipidaemia and hypertension.

The long-term management of patients with diabetes also aims to reduce the risk of the microvascular complications that affect the kidneys, nerves and eyes. Clinicians need to be skilled in strategies that reduce these risks and in the identification and management of complications when they occur.

Optimization of treatment with the growing range of hypoglycaemic agents depends in part on having a good understanding of their mechanisms of action. Advances in the formulation of insulin and production of insulin analogues also offers patients the possibility of improved glycaemic control. Effective management of blood glucose concentrations reduces risks of chronic complications and is particularly important during intercurrent severe illness and pregnancy.

The acute diabetic emergencies of diabetic ketoacidosis and the hyperosmolar hyperglycaemic state are still associated with significant morbidity and mortality. Their management depends on careful adherence of treatment protocols that are based on clear pathophysiological principals. Appropriate, frequent biochemical monitoring of patients with these emergencies is essential to a successful outcome.

ACKNOWLEDGEMENT

The authors wish to acknowledge the contributions of Dr Victor Lawrence and Dr Simon Coppack, the authors of this chapter in the second edition of this book.

Further reading

Gardner DG, Shoback D, editors. Greenspan's basic & clinical endocrinology. 9th ed. New York: McGraw-Hill; 2011.

Holt RIG, Cockram C, Flyvbjerg A et al. editors. Textbook of diabetes: A clinical approach. 4th ed. Oxford: Wiley-Blackwell Scientific; 2010.

A comprehensive account of all aspects of diabetes.

The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med* 1993;329:977–86.

A report of a multicentre, randomized study, which supports the desirability of achieving near-normal plasma glucose concentrations in patients with type 1 diabetes mellitus.

United Kingdom Prospective Diabetes Study. Intensive blood-glucose control with sulfonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet* 1998;352:837–52.

This study provides the evidence for the benefit of tight glycaemic control on the risk of microvascular complications in patients with type 2 diabetes.

Wass JAH, Stewart PM, Amiel SA et al. editors. Oxford textbook of endocrinology & diabetes. 2nd ed. Oxford: Oxford University Press; 2011.

Hypoglycaemia

Mourad H. Labib

CHAPTER OUTLINE

GLUCOSE HOMOEOSTASIS IN THE FED AND THE POSTABSORPTIVE STATES 333

HYPOGLYCAEMIA 334

The neuroendocrine response to hypoglycaemia 334

Symptoms of hypoglycaemia 335

CLASSIFICATION OF HYPOGLYCAEMIC DISORDERS 335

PRACTICAL APPROACH TO THE INVESTIGATION OF HYPOGLYCAEMIA 336

Evaluation of hypoglycaemia in persons without diabetes mellitus 336

Investigation of hypoglycaemia 336

Evaluation of hypoglycaemia in patients with diabetes mellitus 340

EMERGENCY TREATMENT OF HYPOGLYCAEMIA 341

CAUSES OF HYPOGLYCAEMIA 341

Surreptitious administration of hypoglycaemic agents (factitious or felonious hypoglycaemia) 341

Islet cell tumours (insulinoma) 341

Non-insulinoma pancreatogenous hypoglycaemia syndrome (NIPHS) 343

Non-islet cell tumour hypoglycaemia (NICTH) 343

Autoimmune hypoglycaemia 344

Hypoglycaemia associated with renal impairment 344

Hypoglycaemia associated with liver disease 345

Hypoglycaemia due to endocrine deficiencies 345

Drug-induced hypoglycaemia 346

Alcohol-induced hypoglycaemia 346

Hypoglycaemia due to deficient energy intake 347

Septicaemia 347

Exercise-related hypoglycaemia 347

Postprandial (reactive) hypoglycaemia 347

The postprandial syndrome 347

Inherited metabolic disease 348

CONCLUSION 348

GLUCOSE HOMOEOSTASIS IN THE FED AND THE POSTABSORPTIVE STATES

In healthy subjects, the blood glucose concentration is maintained within relatively narrow limits through a tightly controlled balance between glucose production and glucose utilization. Fundamentally, glucose is derived either from dietary intake (in the fed state) or from glycogenolysis and gluconeogenesis (in the fasting or postabsorptive state). It is metabolized by oxidation or stored either in the form of glycogen or through conversion to fat (Fig. 17.1).

Plasma glucose regulation is a complex process involving both insulin-dependent and insulin-independent mechanisms. This regulation is a multi-organ process involving the gastrointestinal tract, pancreas, liver, muscles, adipose tissue, brain and kidneys. In the gastrointestinal tract, incretin hormones (see below), secreted in response to meals, promote glucose-mediated insulin secretion

and suppress glucagon production. In the liver, insulin regulates blood glucose concentrations by suppressing hepatic glucose output and increasing postprandial glucose storage in the form of glycogen. In muscle and adipose tissue, insulin binding to insulin receptors leads to increased expression of glucose transporter 4 (GLUT4) molecules in the cell membrane, facilitating glucose uptake. Insulin-independent mechanisms that contribute to glucose regulation are located in different organs, but mainly in the gastrointestinal tract and the kidneys. Sodium–glucose cotransporters (SGLT), such as SGLT1 in the gastrointestinal tract and SGLT2 in kidney, are important mediators of insulin-independent glucose transport across cell membranes.

In response to nutrient ingestion, glucagon-like peptide (GLP-1) and glucose-dependent insulinotropic peptide (GIP) are released from enteroendocrine cells. Glucagon-like peptide-1 enhances glucose-dependent insulin secretion and inhibits glucagon secretion.

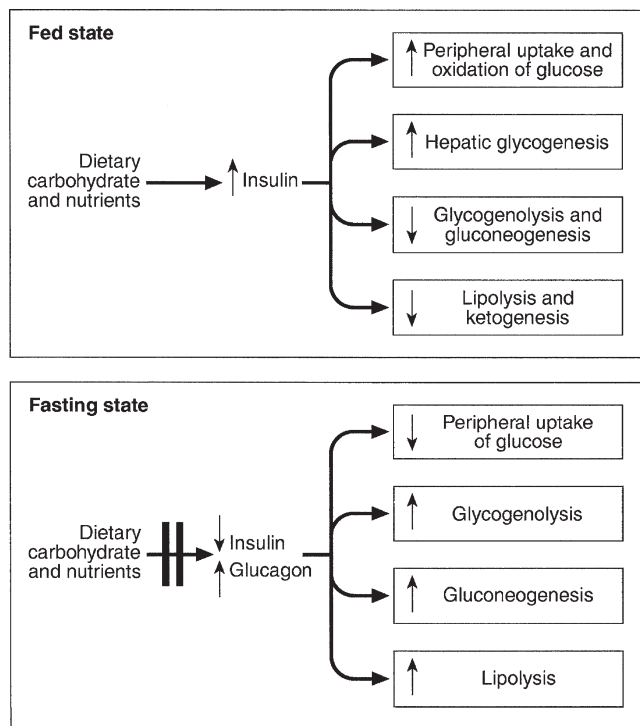


FIGURE 17.1 ■ Glucose homeostasis in the fed and fasting states.

Glucose-dependent insulintropic peptide also stimulates glucose-dependent insulin release, and the combined actions of GLP-1 and GIP account for up to 60% of the overall postprandial insulin secretion. The increase in glucose and insulin concentrations results in stimulation of glycogen synthase and inhibition of glycogen phosphorylase, causing a net increase in hepatic glycogen. Approximately 50% of an oral glucose load is taken up by the liver and 50% by peripheral insulin-sensitive tissues. Three to four hours following a meal, glucose and insulin concentrations decline and the liver switches from net glucose uptake to net glucose release. Hepatic glucose release continues to increase over the next several hours until it equals glucose utilization. Initially, about 75% of glucose production is attributable to glycogenolysis and 25% to gluconeogenesis, but as hepatic glycogen stores decline, the proportion from gluconeogenesis increases steadily.

During the early postabsorptive phase, glucose utilization continues at the rate of 0.10–0.14 mmol/kg per min, of which 40–50% is due to obligatory uptake by the brain and other non-insulin-dependent tissues. With prolonged fasting, the decline in plasma insulin concentration is accompanied by increases in plasma free fatty acid and ketone body concentrations. These can be used as alternative fuels, thereby decreasing the need for glucose.

HYPOGLYCAEMIA

The definition of hypoglycaemia is somewhat arbitrary since the glycaemic threshold at which symptoms occur differs between individuals, and depends on the age of

the patient and the prevailing plasma glucose concentration before the hypoglycaemic episode. A venous plasma glucose concentration <3.0 mmol/L, regardless of the presence or absence of symptoms, has been recommended as the biochemical definition of hypoglycaemia by The Endocrine Society in the USA. However, many apparently healthy people, especially women, may have plasma glucose concentrations <3.0 mmol/L during a prolonged fast without experiencing any symptoms. Conversely, diabetic patients may develop hypoglycaemic symptoms at higher plasma glucose concentrations. Therefore, the diagnosis of pathological hypoglycaemia necessitates the demonstration of Whipple's triad. This consists of: (1) either symptoms or signs, or both, consistent with hypoglycaemia in the presence of (2) a low plasma glucose concentration, with (3) the relief of symptoms or signs after the plasma glucose concentration is raised.

The neuroendocrine response to hypoglycaemia

Under normal metabolic conditions, in the unstressed state, the central nervous system (CNS) is wholly dependent upon glucose as its primary source of energy. The brain requires approximately 150 g of glucose each day, which must be continuously available. The presence of hypoglycaemia triggers a complex and well-coordinated metabolic process aimed at limiting and preventing a further fall in blood glucose concentration. The extent and duration of hypoglycaemia are primary factors in determining the magnitude of this counter-regulatory response, but other factors, such as age, sex, rate of fall of blood glucose and insulin per se have also been shown to influence the response.

In healthy subjects, the initial response to prevent a decline in blood glucose concentration is a reduction in insulin secretion, which begins while the plasma glucose concentration is falling but still within the physiological range. This is followed by an increase in glucagon and adrenaline (epinephrine) secretion as the glucose concentration declines further. Glucagon, secreted by pancreatic α -cells, rapidly raises plasma glucose concentration by stimulating hepatic glucose production via glycogenolysis and gluconeogenesis. The activation of the autonomic nervous system, mainly the sympathetic, increases the concentrations of noradrenaline (norepinephrine) at nerve terminals and adrenaline in the circulation. Adrenaline acts on α - and β -adrenergic receptors at several end organs and results in a more sustained increase in blood glucose concentration. It reduces insulin secretion from the pancreatic islets, increases glycogenolysis and gluconeogenesis in liver, increases glycolysis in muscle and increases lipolysis in adipose tissue. Hepatic glycogen stores provide glucose, particularly for the brain, whereas the mobilization of fatty acids from fat depots provides energy for tissues that can utilize fatty acids as their basic fuel, for example skeletal and cardiac muscles, renal cortex and liver, thus sparing glucose for use by the CNS.

As blood glucose concentration declines further, hypothalamic sensing of hypoglycaemia results in increased secretion of growth hormone (GH) and adrenocorticotrophic hormone (ACTH) and, consequently, cortisol. The increase in GH and cortisol secretion induces metabolic changes over longer periods of time by stimulating lipolysis in adipose tissue and ketogenesis and gluconeogenesis in the liver.

The glycaemic thresholds for the activation of these responses are higher than those for the development of symptoms or the impairment of cognitive function. With falling plasma glucose concentrations, the typical sequence of responses (and their glycaemic thresholds) is as follows: decrease in endogenous insulin secretion (4.4 mmol/L); increase in glucagon and adrenaline (3.8 mmol/L); increase in GH secretion (3.7 mmol/L); increase in cortisol secretion (3.2 mmol/L); development of symptoms (3.0 mmol/L), and impairment of cognitive function (2.5 mmol/L).

Symptoms of hypoglycaemia

The symptoms of hypoglycaemia are non-specific and vary depending on the degree of hypoglycaemia, the rapidity of the decline in blood glucose, the age of the patient, whether the hypoglycaemia is clinical or experimental, and on patients' differing perceptions of symptoms. The symptoms differ from one person to another, but are usually consistent from episode to episode for any one person. In some patients, symptoms are mainly due to activation of the sympathetic nervous system (neurogenic or autonomic), but in others only symptoms due to brain glucose deprivation per se may be observed. All the symptoms of hypoglycaemia, however, reflect the effects of glucose deprivation on the CNS (neuroglycopenia).

Acute neuroglycopenia (neurogenic)

This is characterized by sweating, anxiety, hunger, tremor, palpitations and weakness (Box 17.1). These symptoms result mainly from activation of the sympathetic nervous system and characteristically occur when there has been a rapid decline in blood glucose concentration. It may occur in patients with diabetes owing to excessive absorption of exogenous insulin, either from overtreatment or from rapid mobilization from the injection site during exercise. In non-diabetic subjects, reactive hypersecretion of insulin may be responsible, for example after Roux-en-Y gastric bypass for obesity. Awareness of hypoglycaemia is mainly the result of the perception of neurogenic symptoms which may be lost in patients with type 1 diabetes mellitus (hypoglycaemia unawareness). The symptoms of acute neuroglycopenia may mimic those of anxiety attacks or hyperventilation.

Subacute neuroglycopenia

Symptoms range from a generalized decrease in spontaneous activity, poor concentration, episodic disorientation, confusion, incoordination, slurred speech and behavioural

BOX 17.1 Symptoms of hypoglycaemia (Neuroglycopenia)

Acute	Subacute	Chronic
Sweating	Reduction of spontaneous activity	Personality changes
Anxiety	Behavioural change	Defective memory
Hunger	Tiredness/somnolence	Depression
Nausea	Dizziness	Psychosis
Tremulousness	Poor concentration	Dementia
Weakness	Incoordination	
Palpitations	Headache	
Restlessness	Blurred vision	
	Slurred speech	
	Disorientation/confusion	
	Amnesia	
	Paraesthesia	
	Weakness/transient hemiplegia	
	Seizures	
	Coma	

changes, to seizures, focal neurological signs and coma (Box 17.1). Patients with subacute neuroglycopenic symptoms have been misdiagnosed as having epilepsy, transient ischaemic attacks, psychosis, hysteria, chronic nervous exhaustion, personality disorder or inebriation.

Chronic neuroglycopenia

Chronic neuroglycopenia is very rare and may occur in patients with insulin-secreting tumours unsuspected for years. Symptoms include insidious progressive mental disturbance that may resemble personality disorders, schizophrenia, paranoid psychosis, depression or dementia (Box 17.1).

CLASSIFICATION OF HYPOGLYCAEMIC DISORDERS

Several classifications of hypoglycaemia are possible, for example aetiological, ketotic or non-ketotic, a classification based on the timing of symptoms in relation to meals (fasting or postprandial) or whether hypoglycaemia occurs in a healthy person or one who has an illness.

The usefulness of classifying hypoglycaemic disorders based on the timing of symptoms in relation to meals has been questioned because some conditions that are associated with fasting hypoglycaemia, such as insulinoma, may also produce symptoms postprandially, and post-gastric bypass patients, who typically have postprandial hypoglycaemia, may have symptoms when fasting. Also, patients with factitious hypoglycaemia may have symptoms that have no clear relationship to food intake. In view of this, a classification which is based on the clinical characteristics, i.e. on whether hypoglycaemia occurs in a seemingly well individual or an ill or medicated individual, is more useful (Box 17.2).

BOX 17.2 Clinical classification of hypoglycaemic disorders in adults

III, hospitalized or medicated individual

- Drugs (insulin or insulin secretagogues, alcohol, others)
- Major organ failure (chronic kidney disease, congestive cardiac failure, liver failure)
- Sepsis (including malaria)
- Inanition (anorexia nervosa or starvation)
- Non-islet cell tumour hypoglycaemia (NICTH)
- Hormone deficiencies: cortisol (Addison disease)

Seemingly well individual

- Insulinoma
- Non-insulinoma pancreatogenous hypoglycaemia syndrome (NIPHS)
- Insulin autoimmune hypoglycaemia (antibody to insulin or insulin receptor)
- Factitious or accidental hypoglycaemia (induced by insulin or insulin secretagogues)

PRACTICAL APPROACH TO THE INVESTIGATION OF HYPOGLYCAEMIA

Evaluation of hypoglycaemia in persons without diabetes mellitus

The direction of investigation will depend essentially on the clinical presentation and whether the patient appears ill or not. Hospitalized patients are often severely ill with multisystem disease. In such cases, the underlying illnesses that can produce episodes of hypoglycaemia, such as severe renal or hepatic disease, congestive cardiac failure, sepsis and anorexia nervosa, are usually obvious clinically. A review of all medications is essential since drugs (see below) are a common cause of hypoglycaemia in hospitalized patients, particularly in the presence of renal impairment. Non-islet cell tumours causing hypoglycaemia are usually, though not invariably, large mesenchymal tumours that are clinically apparent. In these cases, confirmation of the suspected mechanism of the hypoglycaemia may be sought by measuring plasma insulin, C-peptide, proinsulin and β -hydroxybutyrate (β -OHB) concentrations during an episode of hypoglycaemia. Endocrine deficiencies such as hypopituitarism and adrenocortical insufficiency should be sought and excluded by appropriate investigations if necessary.

In a seemingly well individual, with no obvious co-existing disease, the direction and extent of evaluation will depend on the clinical presentation. A thorough and detailed history of the symptoms (frequency, type, relationship to meals and exercise) is therefore essential. If symptoms are relieved by food ingestion, it is important to enquire about the type of food and speed of recovery. All medications should be inspected to exclude a prescribing or dispensing error, and the possibility of surreptitiously induced hypoglycaemia should always be considered, especially in healthcare professionals and relatives and carers of diabetic patients. Assessment of

alcohol intake and pattern of drinking is important, since alcohol can cause both fasting and reactive hypoglycaemia. A history of previous gastric surgery or Roux-en-Y gastric bypass surgery for obesity and a family history of multiple endocrine neoplasia type 1 (MEN 1) should be sought. Symptoms in individuals of Japanese or Korean ethnicity may point to autoimmune hypoglycaemia.

Investigation of hypoglycaemia

The aims of the investigation are, first, to demonstrate that hypoglycaemia is the cause of the symptoms and, second, to identify the cause of the hypoglycaemia (Fig. 17.2).

Demonstration of hypoglycaemia

Measurement of blood glucose during spontaneous symptoms. Measurement of the blood glucose concentration (and collection of a suitable blood specimen for subsequent measurement of plasma insulin, C-peptide, proinsulin and β -OHB concentrations) during spontaneous symptoms, and before glucose is given, is without doubt the best test for the diagnosis of spontaneous hypoglycaemia. The most practical way to obtain an accurate glucose concentration is to measure glucose immediately in whole blood or to separate plasma from cells within 30 min of collection, even if the specimen is collected in a tube that contains sodium fluoride. It is well recognized that the rates of decrease of glucose in the first hour after specimen collection in tubes with or without fluoride are virtually identical. The reduction in glucose after 2 h in a fluoride tube can sometimes exceed 0.5 mmol/L leading to a falsely low blood glucose concentration or 'pseudohypoglycaemia'. In addition, although arteriovenous plasma glucose concentration differences are negligible in the fasting state, antecubital venous plasma glucose concentrations are significantly lower than arterial glucose concentrations after a glucose load owing to glucose extraction across the forearm. Therefore, arterialized blood collected from a vein at the back of the hand, which has been warmed by a heat pad, is preferred.

Often, however, patients referred for a medical opinion are asymptomatic when seen in the outpatient clinic, and measuring their blood glucose concentration at such time is usually unhelpful. In this situation, reproducing the circumstances which may lead to hypoglycaemia should be attempted. In patients with a history suggestive of fasting hypoglycaemia, a prolonged fast test should be performed, whereas in patients who only experience symptoms of hypoglycaemia within a few hours of having a meal, a mixed meal test should be performed (see provocation tests, below).

Provocation tests

Prolonged fast test. The prolonged fast (up to 72 h) is the single most useful investigation employed in the evaluation of patients with suspected spontaneous hypoglycaemia. The aim of the investigation is to demonstrate spontaneous hypoglycaemia in the presence of neuroglycopenic symptoms and the resolution of the symptoms when the plasma glucose concentration

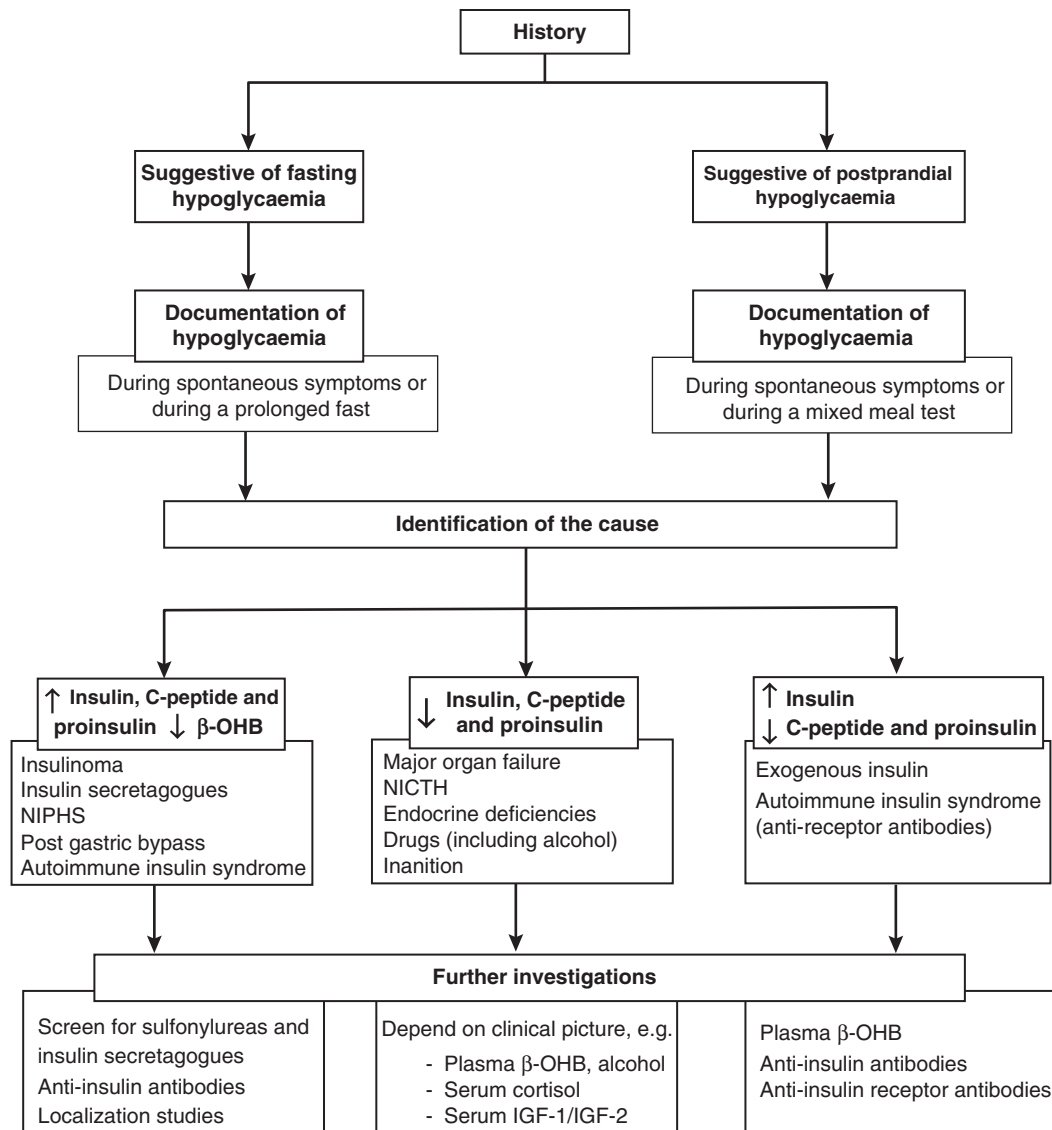


FIGURE 17.2 ■ A practical approach to the investigation and diagnosis of hypoglycaemia (β -OHB, β -hydroxybutyrate; NIPHS, non-insulinoma pancreatogenous hypoglycaemia syndrome; NICTH, non-islet cell tumour hypoglycaemia; IGF-1 and IGF-2, insulin-like growth factors 1 and 2).

is raised to within normal limits. During a prolonged fast, normal subjects rarely develop plasma glucose concentrations <3.0 mmol/L and almost never develop neuroglycopenic symptoms.

The fast must be conducted in hospital under strict medical supervision. Before initiation of the fast, a simple assessment of coordination, recent memory and calculations (e.g. counting down in sevens) should be performed as a baseline. The time for initiating the fast is determined depending on a reasonable estimate of the patient's likely tolerance for fasting. If fasting starts after the evening meal or at midnight, the majority of patients with insulinoma will develop hypoglycaemia by the middle of the following day, when adequate staff and laboratory facilities are available. Patients should have an intravenous indwelling catheter, and an ampoule of 50% glucose solution should be readily available. During the test, the patient is allowed water and non-caloric beverages and should undertake brisk

exercise under supervision. Blood, for glucose, insulin, C-peptide, proinsulin and β -OHB measurements, is collected at the beginning of the test, and every 4–6 h thereafter. Plasma glucose should be determined in the laboratory as soon as possible after collection, and plasma or serum should be stored at -20°C for subsequent analysis of insulin, C-peptide, proinsulin and β -OHB if required. Bedside glucose meters can be used to monitor the patient's blood glucose concentration frequently, but low values, especially in the absence of symptoms of hypoglycaemia, should always be confirmed in the laboratory. When the plasma glucose has fallen to <3.0 mmol/L, blood should be collected more frequently and the patient should be assessed for neuroglycopenia by repeating the same tests performed at the initiation of the fast. If neuroglycopenic symptoms develop and blood glucose is <3.0 mmol/L, several specimens should be obtained for plasma glucose, insulin, C-peptide, proinsulin and β -OHB measurements

before any remedial action, such as glucose administration, is taken. At the end of the fast, a sample should be collected to screen for hypoglycaemic agents and to measure insulin antibodies. A glucagon test (1.0 mg intravenously) has also been recommended (see below). The criteria to stop the test before 72 h are: (1) Whipple's triad has been observed; (2) plasma glucose concentration <3.0 mmol/L in a patient who had previously experienced Whipple's triad, and (3) plasma β -OHB >2.7 mmol/L. The measurement of β -OHB concentrations in real-time during the fast is quite useful because a progressive rise in plasma β -OHB concentrations indicates indirectly that the circulating insulin concentration has been suppressed and that the fast can be terminated. However, blood should always be collected in this situation to confirm suppressed plasma insulin, proinsulin and C-peptide concentrations.

In patients with insulinoma, about 40% develop hypoglycaemia within 12 h, 75% within 24 h, 95% by 48 h and 99% by 72 h.

Glucagon stimulation test. This test serves as a supplemental study in the diagnosis of hypoglycaemic disorders when results from the prolonged fast are inconclusive. The rationale is that, in patients with insulinoma, there is a greater increase in plasma glucose after intravenous glucagon due to the lesser depletion of hepatic glycogen during the fast as a consequence of the higher insulin secretion. The test is performed at the termination of a prolonged fast test, either at the time of occurrence of symptomatic hypoglycaemia with concomitant plasma glucose concentration <3.0 mmol/L, or after 72 h of fast. Glucagon 1 mg is injected intravenously over 2 min and plasma glucose is measured at baseline, 10, 20 and 30 min after the glucagon injection. An increase in plasma glucose concentration of ≥ 1.4 mmol/L after intravenous glucagon indicates mediation of the hypoglycaemia by insulin. A disadvantage of this test is the danger of causing severe hypoglycaemia after 90–180 min.

Mixed meal test. This test is used to investigate patients who experience only postprandial symptoms. The meal has not been standardized but, ideally, a meal similar in composition to the meal that usually provokes symptoms should be used. Plasma glucose concentration is monitored every 30 min for 5 h and also at any time while the patient is symptomatic. Capillary or arterialized blood (see above) should be used, since postprandial venous plasma glucose concentrations are about 10% lower than arterial concentrations owing to extraction of glucose by muscle. Samples should also be collected every 30 min for insulin, C-peptide and proinsulin measurement but should only be analysed if they have been collected at the time when plasma glucose concentration was <3.0 mmol/L and before administering carbohydrates. Patients should be observed for symptoms, and the timing of these symptoms in relation to the meal, and to the blood glucose concentrations, should be documented. If Whipple's triad is demonstrated, a sample for the measurement of oral hypoglycaemic agents and insulin antibodies should be collected. Patients who develop neuroglycopenic symptoms during hypoglycaemia, but not at other times during the test, may be considered to

have postprandial hypoglycaemia. However, in all patients with a positive mixed meal test, additional testing, including a supervised 72 h fast, is recommended. Postprandial hypoglycaemia without fasting hypoglycaemia has been documented in some patients with insulinoma and in patients with non-insulinoma pancreatogenous hypoglycaemia (see p. 343).

The diagnosis of postprandial hypoglycaemia should not be made on the basis of an oral glucose tolerance test (OGTT). At least 10% of 'normal' subjects may have plasma glucose nadirs of <2.6 mmol/L during an OGTT and 2.5% may have values <2.2 mmol/L. Moreover, symptoms during OGTT are often unrelated to the level of plasma glucose nadir or to the rate of decline of the glucose concentration.

Identification of the cause of hypoglycaemia

Plasma insulin, C-peptide and proinsulin. Although insulin and C-peptide are secreted in equimolar amounts by the β -Cells of the pancreas, the metabolic clearance of insulin is much more rapid than that of C-peptide. Therefore, C-peptide has a longer half-life and is present in peripheral blood in higher molar concentrations than insulin, making it less prone to marked fluctuations. Consequently, the measurement of plasma C-peptide concentrations may be more reliable as an indication of endogenous insulin production. However, as C-peptide is cleared by the kidneys, raised concentrations may occur in renal impairment.

Intact proinsulin undergoes enzymatic processing leading to production of des-31,32-proinsulin and des-64,65-proinsulin and then to insulin and C-peptide. Some proinsulin assays specifically measure intact proinsulin whereas others measure 'total proinsulin', i.e. intact and other molecules such as des-31,32-proinsulin. Therefore, proinsulin results depend strongly on the assay used and should be interpreted with caution. In islet cell tumours, the amount of circulating proinsulin is increased and, occasionally, tumours may secrete mainly or exclusively proinsulin. A plasma proinsulin concentration ≥ 5.0 pmol/L, when the plasma glucose concentration is <2.5 mmol/L during a 72 h fast test, represents the best criterion for the diagnosis of endogenous hyperinsulinaemia with 100% specificity and sensitivity. In view of this, the Endocrine Society has now recommended the measurement of plasma proinsulin as a first-line test, in addition to insulin and C-peptide.

Insulin and C-peptide can be measured in either plasma (lithium heparin) or serum, as long as the serum or plasma is separated within 15 min of collection and frozen immediately following separation. Plasma insulin concentrations can be artefactually modified by heterophilic antibodies, endogenous anti-insulin antibodies or haemolysis, as red blood cells contain an insulin-degrading enzyme, which may lead to underestimation of the actual insulin concentrations in haemolysed samples.

The measurement of plasma insulin, C-peptide and proinsulin concentrations, in the presence of hypoglycaemia, is the most useful test in identifying the cause

of hypoglycaemia (see Fig. 17.2). Raised plasma insulin (≥ 18 pmol/L), C-peptide (≥ 200 pmol/L) and proinsulin concentrations (≥ 5.0 pmol/L) in the presence of hypoglycaemia (plasma glucose < 3.0 mmol/L) indicate endogenous hyperinsulinaemia. Causes of endogenous hyperinsulinaemia include insulinoma, non-insulinoma pancreatogenous hypoglycaemia syndrome (NIPHS), autoimmune insulin syndrome, sulfonylurea- or meglitinide-induced hypoglycaemia, non-islet cell insulin-secreting tumours and reactive insulin-mediated hypoglycaemia in post-gastric bypass surgery patients. Factitious sulfonylurea-induced and meglitinide- (glinide-) induced hypoglycaemia may produce an identical clinical and biochemical picture to insulinoma. Therefore, screening for these drugs is essential if a false positive diagnosis of insulinoma is to be avoided.

Inappropriately raised plasma insulin concentrations in the presence of low or suppressed plasma C-peptide and proinsulin concentrations will identify patients with exogenous insulin administration. Therefore, it is important to use insulin assays that can detect the presence of insulin analogues. Another rare cause of raised plasma insulin but suppressed C-peptide and proinsulin concentrations is hypoglycaemia mediated by insulin receptor antibodies (IR-A). The diagnosis should be considered in patients with other autoimmune diseases and requires the demonstration of IR-A in the serum.

Other causes of hypoglycaemia are associated with suppressed plasma insulin, C-peptide and proinsulin concentrations. In many of these conditions, the diagnosis is usually obvious on clinical grounds. In others, the measurement of plasma β -OHB, GH and insulin-like growth factor (IGF-1 and IGF-2) concentrations may be required to establish the cause of hypoglycaemia (Fig. 17.2).

Plasma β -hydroxybutyrate (β -OHB). Lipolysis is very sensitive to circulating insulin concentrations. During fasting, normal individuals will show a gradual decline in insulin concentration and a progressive increase in lipolysis and hence ketone bodies, for example β -OHB (Fig. 17.3). In patients with hypoglycaemia due to hyperinsulinaemia, lipolysis is suppressed and β -OHB concentrations are low. A plasma β -OHB concentration of ≤ 2.7 mmol/L during a prolonged fast, and when plasma glucose concentration is < 3.0 mmol/L, indicates mediation of hypoglycaemia by insulin. Conversely, in hypoglycaemia associated with suppressed insulin secretion, such as liver disease, β -OHB concentrations are usually raised, the only exception being in severe inanition, when fat stores are depleted. Therefore, the measurement of plasma β -OHB provides an indirect measure of the prevailing insulin concentration during the hypoglycaemia. The main advantage, however, is that β -OHB is easy to measure and the result can be made available long before those of insulin or C-peptide measurements.

Insulin antibodies. The presence of circulating antibodies to insulin, due to previous exposure to exogenous insulin, may give falsely high plasma insulin concentrations (see p. 295). However, this is less common now with

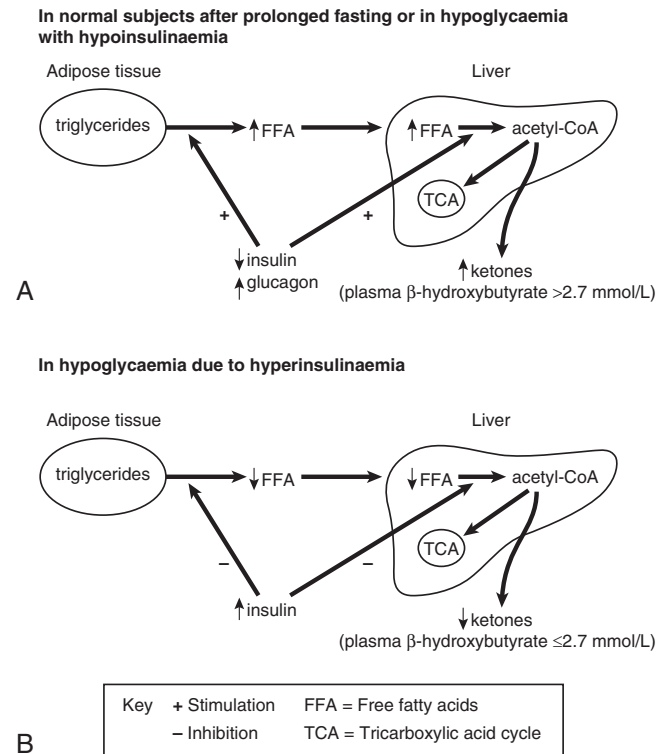


FIGURE 17.3 ■ Changes in the concentrations of hormones and metabolites in normal subjects after prolonged fasting and in hypoglycaemia with hypoinsulinaemia (A), and in hypoglycaemia due to hyperinsulinaemia (B).

the use of human insulin, which is less antigenic than the previously used animal insulin preparations. The detection of insulin antibodies is considered to be the criterion for the diagnosis of the insulin autoimmune syndrome (see below), but antibodies may be detected in persons without hypoglycaemia and even, rarely, in patients with insulinoma. Because C-peptide does not cross-react with insulin antibodies, the measurement of C-peptide in these situations can be used as an index of β -Cell function. Autoantibodies to insulin receptors cause a particularly refractory type of fasting hypoglycaemia (see below) and the diagnosis requires the demonstration of IR-A in the serum.

Screening for oral hypoglycaemic agents. In a seemingly well individual, screening for oral hypoglycaemic agents (sulfonylureas and meglitinides) should always be considered to exclude the possibility of a prescribing or dispensing error, and of surreptitiously induced hypoglycaemia.

Evaluation of hypoglycaemia in patients with diabetes mellitus

Definition

The American Diabetes Association (ADA) has published criteria for the definition and clinical classification of hypoglycaemia in patients with diabetes mellitus (Box 17.3). In patients with diabetes, hypoglycaemia is defined as all episodes of an abnormally low plasma

BOX 17.3 American Diabetes Association classification of hypoglycaemia in individuals with diabetes**Severe hypoglycaemia**

- An event requiring assistance of another person to actively administer carbohydrate, glucagon or other resuscitative actions. These episodes may be associated with sufficient neuroglycopenia to induce seizure or coma. Plasma glucose measurements may not be available during such an event, but neurological recovery attributable to the restoration of plasma glucose to normal is considered sufficient evidence that the event was induced by a low plasma glucose concentration.

Documented symptomatic hypoglycaemia

- An event during which typical symptoms of hypoglycaemia are accompanied by a measured plasma glucose concentration ≤ 3.9 mmol/L.

Asymptomatic hypoglycaemia

- An event not accompanied by typical symptoms of hypoglycaemia but with a measured plasma glucose concentration ≤ 3.9 mmol/L.

Probable symptomatic hypoglycaemia

- An event during which symptoms of hypoglycaemia are not accompanied by a plasma glucose determination (but that was presumably caused by a plasma glucose concentration ≤ 3.9 mmol/L).

Relative hypoglycaemia

- An event during which the person with diabetes reports any of the typical symptoms of hypoglycaemia and interprets those as indicative of hypoglycaemia, but with a measured plasma glucose concentration > 3.9 mmol/L.

glucose concentration (with or without symptoms) that expose the individual to harm. The workgroup recommended that persons with diabetes become concerned about the possibility of hypoglycaemia at a self-monitored blood glucose concentration ≤ 3.9 mmol/L. While that value is higher than the value used to diagnose hypoglycaemia in people without diabetes, it approximates the lower limit of the physiological fasting non-diabetic range and the glycaemic threshold for activation of counter-regulatory mechanisms. This cut-off value has been debated, with some favouring a value of < 3.5 mmol/L to avoid over-diagnosis of hypoglycaemia in asymptomatic patients.

Pathophysiology and risk factors

Hypoglycaemia in persons with diabetes is essentially the result of therapies that raise plasma insulin concentrations, such as insulin, sulfonylureas and non-sulfonylurea insulin secretagogues (e.g. nateglinide or repaglinide). The relative or absolute insulin excess during treatment, together with compromised physiological and behavioural defences in type 1 diabetes and longstanding type 2 diabetes, can result in hypoglycaemic episodes. In many patients with fully developed type 1 diabetes or with longstanding type 2 diabetes, the glucagon response to hypoglycaemia is lost and the adrenaline response is often reduced, leading to defective counter-regulation and 'hypoglycaemia unawareness'.

Factors that may increase risk of hypoglycaemia include: missed or irregular meals; incorrect use of hypoglycaemic medication (dose or timing); decreased insulin clearance (e.g. renal impairment); increased glucose utilization (e.g. excessive exercise); excessive alcohol intake, and concurrent medications (Box 17.4).

Incidence

The rates of hypoglycaemia with insulin therapy vary according to the insulin regimen and the type and duration of diabetes. Overall, hypoglycaemia is less frequent in type 2 diabetes than in type 1 diabetes, but the risk of hypoglycaemia increases substantially later in the course

BOX 17.4 Risk factors for hypoglycaemia in diabetes

- Excessive or incorrect use of insulin or insulin secretagogues (e.g. dose or timing)
- Impaired drug clearance (e.g. renal impairment, liver failure, hypothyroidism)
- Impaired counter-regulatory capacity (e.g. GH deficiency, Addison disease)
- Decreased exogenous glucose delivery (e.g. missing meals, anorexia)
- Impaired glucose absorption (e.g. malabsorption)
- Decreased endogenous glucose production (e.g. liver failure, alcohol)
- Increased peripheral glucose utilization (e.g. exercise)
- History of severe hypoglycaemia, hypoglycaemia unawareness or recent antecedent hypoglycaemia
- Concurrent medications (e.g. aspirin, warfarin)

of type 2 diabetes. According to the UK Hypoglycaemia Study Group (2007), the rate of severe hypoglycaemia in insulin-treated type 1 patients of < 5 years' duration was 22%, rising to 46% in patients with duration of > 15 years. In comparison, the rates in patients with type 2 diabetes treated with insulin were 7% for patients taking insulin for < 2 years and 25% for patients taking insulin for > 5 years. Diabetic patients who are treated with insulin are particularly prone to exercise-induced hypoglycaemia. In normal subjects, glucose uptake in skeletal muscles is increased by 20–30 times during exercise. This is compensated for by an increase in hepatic glucose release mediated by a fall in circulating insulin concentrations. In patients treated with insulin, the continuous release of insulin from subcutaneous depots inhibits glucose output from the liver. In addition, increased absorption of insulin from the injection site may occur if the site is near the muscles being exercised.

Hypoglycaemia rates with the third-generation sulfonylureas (e.g. glimepiride, glipizide and gliclazide) and the meglitinides (e.g. repaglinide and nateglinide) appear to be lower than those of previous generation sulfonylureas (e.g. glibenclamide and chlorpropamide). Elderly diabetic patients, especially those with renal or liver impairment, are particularly susceptible to hypoglycaemia

and a reduction in sulfonylurea dosages is often required. The hypoglycaemia is usually prolonged, especially as the duration of action of the sulfonylureas is generally much longer than their plasma half-lives, and may last between 36 h and seven days, despite continuous infusion of glucose. Additional treatment with glucagon and/or diazoxide may be required to prevent a fatal outcome.

The risk of hypoglycaemia with metformin and thiazolidinediones (e.g. pioglitazone) is negligible. These agents act by sensitizing peripheral tissues to insulin but do not directly alter insulin secretion. They rarely cause hypoglycaemia when used as monotherapy, but may contribute to hypoglycaemia when used concomitantly with insulin, sulfonylureas or other insulin secretagogues.

Glucagon-like peptide-1 analogues (such as exenatide or liraglutide) and the dipeptidyl-peptidase 4 inhibitors (such as sitagliptin, vildagliptin and saxagliptin) do not appear to be associated with increased risk of hypoglycaemia when used as monotherapy, or in combination with insulin sensitizers (such as metformin and thiazolidinediones) in short-term clinical studies, but the risk of hypoglycaemia appears to increase when GLP-1 analogues are used with sulfonylureas.

Management

In patients with documented symptomatic or asymptomatic hypoglycaemic events, excessive dosing, wrong type or ill-timed dosing of insulin or insulin secretagogues should be first considered. Conditions that may affect exogenous glucose delivery (e.g. missed meals), endogenous glucose production (e.g. alcohol), glucose utilization (e.g. exercise), insulin sensitivity (e.g. weight loss) or clearance (e.g. renal impairment) should be explored. Recurrent hypoglycaemic episodes or a history of hypoglycaemia unawareness should prompt a detailed review of the treatment regimen.

EMERGENCY TREATMENT OF HYPOGLYCAEMIA

Hypoglycaemia should be suspected in all patients presenting to the emergency department with altered mental state. In these patients, the diagnosis of hypoglycaemia should be expeditiously investigated by bedside glucose measurement; blood sampling for laboratory glucose testing should always be performed to confirm the results, though therapy should never be withheld pending laboratory confirmation. If hypoglycaemia is present, glucose should be administered immediately, preferably orally if the patient is awake and can swallow water safely with no risk of aspiration. The dose of oral glucose should be 10–20 g in an adult patient, and this dose may be repeated three times (if required) before giving a meal containing complex carbohydrates, to ensure long-lasting effects.

If the patient's conscious level does not allow oral therapy, parenteral glucose is indicated: the dose of glucose for an average adult is 25–50 g given as an i.v. bolus of 50–100 mL of 50% dextrose. This should be followed by a maintenance dose of 10% dextrose i.v. at 3 mL/kg per h.

Failure to respond to glucose is an indication to give glucagon subcutaneously, intramuscularly or intravenously (1 mg in adults). Steroids (e.g. hydrocortisone 1–2 mg/kg i.v. every 6 h) are given if the patient remains refractory to standard hypoglycaemia therapy or if adrenal insufficiency is suspected. Sulfonylureas are an important cause of refractory hypoglycaemia. They act by stimulating insulin secretion, and the administration of glucose to treat the hypoglycaemia further increases insulin release, thus tending to perpetuate the hypoglycaemia. Therefore, diazoxide or octreotide may be indicated for patients with sulfonylurea-induced hypoglycaemia.

CAUSES OF HYPOGLYCAEMIA

Surreptitious administration of hypoglycaemic agents (factitious or felonious hypoglycaemia)

Surreptitious or self-induced factitious hypoglycaemia should always be considered in patients presenting with spontaneous hypoglycaemia. The diagnosis should be suspected in individuals with access to insulin, sulfonylureas or meglitinides, such as diabetic patients (and their relatives, friends and carers) and healthcare professionals. Patients often vehemently deny self-administration and rarely show signs of psychiatric disturbances. Patients with diabetes may claim to have hypoglycaemic attacks despite omitting their insulin or sulfonylurea drugs. Some have the characteristics of Munchausen syndrome, with a history of frequent hospitalizations, previous surgical procedures and extensive travel.

Once suspected, the patient should be admitted to hospital for detailed observation and blood should be collected during hypoglycaemia for the measurement of plasma insulin, C-peptide, proinsulin, insulin antibodies and oral hypoglycaemic agents. A raised plasma insulin concentration and a suppressed C-peptide concentration will confirm the diagnosis of self-administration of insulin. If plasma insulin and C-peptide concentrations are both inappropriately raised, a drug screen (ideally for all available sulfonylureas and meglitinides) should be performed, in order to avoid an incorrect diagnosis of insulinoma. Once the diagnosis of factitious hypoglycaemia is made, the patient should be referred for counselling and psychiatric therapy.

Accidental intake of sulfonylurea drugs or meglitinides is rare but should be considered in all patients presenting with spontaneous hypoglycaemia. The finding that a spouse or a relation is taking these drugs should alert the physician to the possibility of accidental intake. The possibility of a prescription or dispensing error should also be considered.

Islet cell tumours (insulinoma)

Insulinomas are rare, with an annual incidence of about four cases per million persons. They can occur at any age, with approximately 80% occurring between the ages of 20 and 60 years and with, probably, a slight preponderance in women. In approximately 80% of

patients, there is a solitary benign adenoma. Benign adenomas are generally small (1–2 cm in diameter) and occur with equal frequency in the head, body and tail of the pancreas. Multiple tumours occur in about 10% of patients and approximately 10% of insulinomas are malignant, with regional or distant metastases. About 8% of insulinomas are associated with multiple endocrine neoplasia type 1 (MEN1). Diffuse islet cell hyperplasia and non-insulinoma pancreatogenous hypoglycaemia syndrome (see below) have been reported in adults. Many insulinomas contain and secrete other hormones in addition to insulin (glucagon, gastrin, somatostatin and pancreatic polypeptide), but hypersecretion of these hormones is not usually associated with clinical symptoms.

Clinical features

Hypoglycaemia is the clinical hallmark of insulinomas. Neuroglycopenic symptoms, especially confusion, weakness, slurred speech, blurred vision and irrational behaviour, typically occur before breakfast, though 50% of patients may also have symptoms late in the afternoon. More than 50% of patients have amnesia and therefore the history should always be corroborated. Nocturnal hypoglycaemia is not uncommon and can manifest as nightmares and morning headaches. Symptoms may be provoked by exercise, missing a meal or by following a restricted calorie diet. Transient hemiplegia, grand mal seizures, focal neurological signs and coma can occur, but permanent brain damage is rare and is only seen with profound hypoglycaemia that goes unrecognized and uncorrected for several hours.

Diagnosis

Insulinoma is a rare cause of spontaneous hypoglycaemia. More common causes, such as surreptitious or inadvertent administration of insulin or oral hypoglycaemic agents, should be excluded first. The diagnosis of insulinoma requires the demonstration of hypoglycaemia, unsuppressed or elevated plasma insulin concentrations during hypoglycaemia and a pancreatic tumour. Patients with suspected insulinoma should have a prolonged fast with monitoring of plasma glucose, insulin, C-peptide, proinsulin and β -OHB concentrations followed by a glucagon stimulation test (if indicated). In patients with insulinoma, plasma insulin, C-peptide and proinsulin concentrations fail to suppress despite the presence of hypoglycaemia. Diagnostic cut-off values in the presence of hypoglycaemia are plasma insulin ≥ 18.0 pmol/L, C-peptide ≥ 200 pmol/L, proinsulin ≥ 5.0 pmol/L, β -OHB concentrations ≤ 2.7 mmol/L and an increase in plasma glucose concentration of ≥ 1.4 mmol/L after intravenous glucagon. The plasma insulin/C-peptide molar ratio is of the order of 1:5, compared with approximately 1:10 in normal subjects, owing to decreased hepatic extraction of insulin. Plasma proinsulin concentrations are significantly higher in patients with insulinoma than in normal subjects, and the molar ratio of insulin/proinsulin is about 1:1 compared with approximately 6:1 in normal subjects.

Localization

Preoperative localization should only be attempted if a definitive biochemical diagnosis has been established. The mean diameter of insulinomas is usually < 2.0 cm, and non-invasive imaging techniques may fail to localize them. Computerized tomography (CT) detects 70–80% and magnetic resonance imaging (MRI) scanning about 80% of insulinomas. A substantial percentage of insulinomas express cell surface receptors for somatostatin, and somatostatin receptor scintigraphy has been reported to have a sensitivity of 80%. However, false localization is not uncommon with non-invasive modalities (i.e. the location suggested by these procedures may not correspond to the tumour found at operation). Invasive imaging procedures, such as endoscopic pancreatic ultrasonography, selective coeliac axis angiography and trans-hepatic portal venous sampling, have been reported to have sensitivities of $> 90\%$. Selective intra-arterial calcium stimulation with hepatic venous sampling (arterial stimulation and venous sampling, ASVS), using plasma C-peptide and insulin gradients, has been shown to have a high sensitivity of 94% and is used when other imaging procedures are equivocal or negative. However, it is the procedure of choice for confirming non-insulinoma pancreatogenous hypoglycaemia syndrome and post-Roux-en-Y gastric bypass hypoglycaemia, as other imaging techniques are negative in those disorders.

Treatment

Surgical resection is the treatment of choice for insulin-secreting tumours and is curative in over 90% of patients. In about 5–10% of patients, no tumour can be found at laparotomy. In these cases, the patient needs to be sent to a centre of excellence for further imaging and re-operation. The previous practice of a blind distal two-thirds pancreatectomy is no longer recommended. In some centres, the measurement of intraoperative plasma insulin concentrations has been used to ensure complete tumour removal. This may be important in patients with MEN1, who may harbour multiple insulinomas. The recurrence rate after surgical resection is 7% but is much higher, at 21%, for those with MEN1.

If surgery is contraindicated, medical treatment using diazoxide or a somatostatin analogue, such as octreotide, can reduce insulin release. Diazoxide, a non-diuretic benzothiazide, suppresses insulin release in about 50% of patients. It can also be useful, in responsive cases, in preoperative management. It directly inhibits insulin release via its effect on the K^+ -ATP channel. The standard dose is 150–450 mg daily but may be increased to as much as 800 mg/day. Its main side-effect is sodium and water retention, which can be prevented by giving a thiazide diuretic. About 50% of patients with insulinoma benefit from octreotide. The usual dose is 100 μ g injected three times daily, and side-effects, such as diarrhoea and steatorrhoea, are rare. However, the effect of octreotide is dependent on the presence of somatostatin receptor subtype 2 on insulinoma tumour cells. A positive OctreoScanTM is not a prerequisite before starting octreotide treatment

since studies have shown that, in patients with insulinoma and a negative scan, somatostatin decreased insulin concentrations significantly and reduced the incidence of hypoglycaemic episodes. In malignant insulinoma, cytotoxic chemotherapy with streptozotocin, doxorubicin and 5-fluorouracil can be used. When systemic chemotherapy is unsuccessful in patients with unresectable metastatic disease to the liver, embolization of the hepatic artery and intra-arterial chemotherapy may be used to control symptoms, to inhibit tumour growth and to improve survival.

Non-insulinoma pancreatogenous hypoglycaemia syndrome (NIPHS)

A rare syndrome of endogenous hyperinsulinaemic hypoglycaemia has been described in adult patients who suffer from postprandial insulin-mediated hypoglycaemia. The syndrome is characterized by neuroglycopenic episodes within 4 h of meal ingestion, negative prolonged fast tests, negative perioperative localization studies for insulinoma but positive intra-arterial calcium stimulation tests. Histologically, the pancreas shows diffuse islet hypertrophy, sometimes with hyperplasia, and enlarged and hyperchromatic β -Cell nuclei (nesidioblastosis). Patients do not have mutations of the *SUR1* (*ABCC8*) or *KIR6.2* (*KCNJ11*) genes, which encode the subunits of the pancreatic ATP-sensitive potassium channel responsible for glucose-induced insulin secretion, and have been described as a cause of hyperinsulinaemic hypoglycaemia in children. Therefore, it has been proposed that adult patients with endogenous hyperinsulinaemic hypoglycaemia may have mutations at other genes, as yet unidentified, that are involved in the regulation of insulin secretion.

Some patients who have undergone Roux-en-Y gastric bypass surgery for obesity develop postprandial neuroglycopenic symptoms, as a result of endogenous hyperinsulinaemia. These symptoms are different from those ascribed to the dumping syndrome, which are characterized by sweating, dizziness and flushing, but not neuroglycopenia. The precise mechanisms of hypoglycaemia in these patients have not yet been determined but increased secretion of GLP-1 and decreased ghrelin concentrations have been implicated. Nesidioblastosis was reported in some patients but not in others and, therefore, whether gastric bypass-induced hypoglycaemia is the result of a structural pancreatic abnormality or a functional dysregulation of postprandial insulin secretion remains to be established. Some patients respond to dietary measures or medications (α -glucosidase inhibitor, diazoxide, calcium channel blockers, octreotide) but others require partial pancreatectomy.

Non-islet cell tumour hypoglycaemia (NICTH)

Non-islet cell tumour hypoglycaemia (NICTH) is the term applied to hypoglycaemia caused by tumours other than insulinomas. A wide variety of tumours have been associated with hypoglycaemia (Box 17.5). Most patients are elderly and present with predominantly neuroglycopenic symptoms that may antedate the recognition of the

BOX 17.5 Non-islet cell tumours associated with hypoglycaemia

Tumours of the connective tissues (mesodermal)

- Fibroblastoma
- Fibroma
- Fibrosarcoma
- Haemangioendothelioma
- Histiocytoma
- Leiomyosarcoma
- Liposarcoma
- Mesenchymal tumours of the diaphragm, pancreas and pleura
- Mesotheliomas
- Rhabdomyosarcoma

Tumours of epithelial origin

- Adrenal carcinoma
- Carcinoma of the breast and cervix
- Carcinoma of the gastrointestinal tract and lung
- Carcinoma of the prostate
- Hepatoma
- Pheochromocytoma
- Wilms tumour and hypernephroma

Tumours of haemopoietic origin

- Leukaemia (chronic myelocytic)
- Lymphoma
- Myeloma

Apudomas

- Carcinoid tumours

existence of the tumour by many years. In most cases, especially those due to large mesenchymal tumours, the tumour is readily detectable by physical or radiological examination, but in others the tumour may be small and difficult to detect.

The pathogenesis of hypoglycaemia has been related to the secretion of incompletely processed insulin-like growth factor 2 (pro-IGF-2) by these tumours. This form of IGF-2 binds poorly to its binding proteins and penetrates more freely into tissue spaces. Normally, both IGF-1 and IGF-2 circulate almost completely bound to specific IGF-binding proteins (IGFBPs). They can be found as binary complexes (IGF-1 or IGF-2 with IGFBP-3) or ternary complexes (IGF-1 or IGF-2 with IGFBP-3 and an acid-labile subunit, ALS). In normal human plasma, 70–80% of IGF-2 is associated with a ternary 150 kD IGFBP complex, consisting of IGF-2, IGFBP-3 and an ALS, while 20–30% is bound to a binary complex of <60 kD consisting of IGF-2 and IGFBP-3, and <2% is present in the free form. Ternary complexes appear to be confined to the circulation, whereas binary complexes appear to leave the circulation almost as rapidly as the free form. In NICTH, IGFBP-3 and ALS production is decreased and the ability of the binary complex to form the more stable ternary complex is impaired; thus tumour-derived IGF-2 appears mainly as binary IGF-IGFBP complexes, which have greatly increased turnover compared with ternary complexes. Although circulating concentrations of IGF-2 are usually normal, the ratio of pro-IGF-2 to IGF-2 is raised. Pro-IGF-2 interacts with insulin receptors and

IGF-1 receptors and inhibits GH secretion, which in turn leads to a decreased plasma IGF-1 concentration. The low GH concentrations also result in a reduction in IGFBP-3, contributing to the greater bioavailability of pro-IGF-2. The hypoglycaemia of tumours secreting pro-IGF-2 is therefore characterized by suppressed insulin and C-peptide, inappropriately low GH and β -OHB concentrations, and the molar ratio of total plasma IGF-2/IGF-1 is usually >10 .

Surgical removal of a benign or locally invasive tumour usually results in complete remission of hypoglycaemia. Treatment with human GH (hGH) and/or prednisolone has also been used with good effect. Growth hormone is thought to exert its beneficial effect by stimulating gluconeogenesis and hepatic glycogenolysis, and by increasing IGFBP-3 production and redistribution of plasma IGFBP-3 from binary to ternary form, potentially reducing the bioavailability of pro-IGF-2. However, this effect is induced only at supraphysiological GH doses. Prednisolone, at daily doses of 30 mg or more, alleviates hypoglycaemia by direct stimulation of gluconeogenesis and glycogenolysis, and may also suppress tumour production of pro-IGF-2. Although both treatments are effective in relieving hypoglycaemia, prednisolone may have greater long-term benefit through suppression of pro-IGF-2.

Autoimmune hypoglycaemia

Two autoimmune syndromes that can cause hypoglycaemia have been described. In one of these, antibodies bind insulin receptors and mimic the action of insulin. The hypoglycaemia may be either fasting or postprandial and is often severe. Some patients may have a preceding phase of hyperglycaemia and significant insulin resistance, whereas in other patients, the onset of hypoglycaemia may be the first manifestation of the syndrome. Most patients with this syndrome also have evidence, either in the laboratory (elevated titres of antinuclear, antithyroid, antimitochondrial or antiplatelet antibodies) or clinically, of other autoimmune diseases such as systemic lupus erythematosus, primary biliary cirrhosis or Hashimoto thyroiditis (Box 17.6). Laboratory investigations show high plasma insulin concentrations (as a result of interference with the insulin assay) but suppressed C-peptide and proinsulin concentrations, similar to the findings with surreptitious administration of insulin. Therefore, the demonstration of the presence of antibodies directed against the insulin receptor (IR-A) is essential to make the diagnosis. The hypoglycaemia seems to respond rapidly to high doses of steroids but not to immunosuppression or plasmapheresis. Prognosis is poor, but in those who survive, the antireceptor antibodies disappear and remission occurs over several months or years.

The other syndrome, the autoimmune insulin syndrome, in which antibodies are directed towards insulin, has been reported mainly in Japan and is extremely rare. Most patients are middle-aged, with no history of previous administration of insulin, and may have evidence of other autoimmune diseases such as Graves disease, rheumatoid arthritis or systemic lupus erythematosus (see Box 17.6). An association with administration of certain drugs such as carbimazole, hydralazine, procainamide, penicillamine,

BOX 17.6 Conditions associated with autoimmune hypoglycaemia

Conditions associated with insulin receptor antibodies

Autoimmune diseases

- Alopecia
- Autoimmune thrombocytopenia
- Hashimoto thyroiditis
- Primary biliary cirrhosis
- Systemic lupus erythematosus
- Sjögren syndrome

Malignancy

- Hodgkin disease

Conditions associated with insulin antibodies

Autoimmune diseases

- Graves disease
- Rheumatoid arthritis
- Systemic lupus erythematosus
- Polymyositis

Medications

- Insulin (human, animal)
- Drugs containing sulfhydryl or thiol groups (captopril, carbimazole, D-penicillamine, gold thioglucose, glutathione)
- Hydralazine
- α -Interferon
- Isoniazid
- Procainamide

Others

- Alcoholic liver disease
- Benign monoclonal gammopathy
- Multiple myeloma
- Post-pancreatic transplant

glutathione and captopril has also been reported. The hypoglycaemia is most often postprandial, although fasting hypoglycaemia has been reported. Presentation can often be dramatic owing to severe neuroglycopenia with confusion and even coma. Paradoxically, hyperglycaemia may occur immediately following a meal or oral glucose challenge. Several possible mechanisms for the hypoglycaemia have been suggested. Sudden dissociation of insulin from the antibodies is the most plausible one. Laboratory investigations reveal high plasma insulin concentrations but, because the insulin is produced endogenously, C-peptide concentrations are not completely suppressed. The majority of circulating insulin is bound to antibodies, and the demonstration of the presence of anti-insulin antibodies is essential to make the diagnosis. In most cases, the hypoglycaemia is usually transient and resolves spontaneously. In severe cases, dietary treatment with frequent low-carbohydrate meals and acarbose has produced some benefit in alleviating the hypoglycaemia.

Hypoglycaemia associated with renal impairment

Renal impairment is a common predisposing factor for hypoglycaemia and is probably the second most common

cause of hypoglycaemia, after insulin therapy, in hospitalized patients. The most important predisposing factor for hypoglycaemia in chronic kidney diseases (CKD) is caloric restriction, whether acute, due to anorexia and vomiting, or chronic. Other predisposing factors include concomitant liver disease, congestive heart failure, sepsis and drug therapy such as insulin, oral hypoglycaemic agents and β -blockers. The pathogenesis of hypoglycaemia in renal failure is complex and several mechanisms have been proposed. In normal subjects, the kidneys play a major role in gluconeogenesis and may supply as much as 45% of glucose during prolonged starvation. In uraemic patients, who are often malnourished, renal gluconeogenesis may not be able to maintain an adequate glucose supply, even if hepatic gluconeogenesis is normal. Other mechanisms include impaired hepatic glycogenolysis and gluconeogenesis; increased insulin half-life due to decreased renal degradation; diminished availability of alanine, and impaired counter-regulatory mechanisms.

Therapy for diabetes mellitus, with insulin or oral hypoglycaemic drugs, is by far the most common cause of hypoglycaemia in patients with renal impairment. Decreased renal degradation of insulin, frequently leading to a reduction in insulin or sulfonylurea requirements, is the main reason. Iatrogenic hypoglycaemia may be a consequence of haemodialysis or peritoneal dialysis. In patients undergoing haemodialysis, the mechanism may be due, in part, to post-dialysis glucose-induced hyperinsulinaemia caused by high glucose content of the dialysate and partly due to impaired insulin degradation by the kidneys.

The symptoms of uraemic hypoglycaemia are mainly neuroglycopenic (drowsiness, headache, lethargy, confusion, convulsions and coma) and may be confused with the symptoms of the dialysis dysequilibrium syndrome. Spontaneous hypoglycaemia in renal failure, not associated with hypoglycaemic agents or dialysis, carries a grave prognosis, with over half of the patients dying within months of onset.

Since the clearance of insulin, C-peptide and proinsulin is impaired in CKD, these parameters are unreliable in excluding the diagnosis of insulinoma in such patients. In these situations, β -OHB, measured during a prolonged fast test, and the glucose response to glucagon have been reported to be the best diagnostic parameters.

Hypoglycaemia associated with liver disease

Although the liver plays a central role in normal glucose homeostasis, hypoglycaemia is rare in patients with liver disease. Glucose homeostasis can be maintained with the mass of functioning hepatocytes reduced to <20% of normal, and hypoglycaemia does not occur until the liver is extensively damaged. Hypoglycaemia is not usually a feature of chronic hepatic failure nor is it implicated in the production of hepatic coma. Conversely, it may occasionally be associated with mild liver disease, and has been reported in a wide variety of liver disease such as fatty infiltration, portal cirrhosis, infective hepatitis and hepatocellular carcinoma.

Frank hypoglycaemia is also uncommon in acute (fulminant) hepatic failure, whether due to viral infection or

hepatotoxic drugs. When it does occur, it can be severe and persistent. The diagnosis of hepatogenous hypoglycaemia may be difficult since the extent of liver disease, as assessed by standard liver function tests, does not correlate with the degree of hypoglycaemia. Therefore, other causes of hypoglycaemia should always be sought in a patient with hypoglycaemia and abnormal liver function tests. The hypoglycaemia associated with congestive heart failure, sepsis and Reye syndrome is thought to be due to hepatic mechanisms.

Hypoglycaemia due to endocrine deficiencies

Endocrine deficiencies, especially those involving the pituitary and the adrenal glands, can result in low blood glucose concentrations but these are rarely <3.0 mmol/L unless associated with other disorders. Spontaneous hypoglycaemia due to pituitary insufficiency is more commonly seen in neonates and children than in adults but may occasionally be the presenting feature, especially in the elderly. The hypoglycaemia is mainly due to GH deficiency, but the associated deficiency of ACTH accentuates the tendency to hypoglycaemia. Spontaneous hypoglycaemia is also recognized in isolated ACTH deficiency and isolated GH deficiency, particularly after prolonged fasting. The diagnosis is usually suspected on clinical examination and is confirmed by appropriate combined pituitary function tests (see Chapter 18). Plasma insulin and C-peptide concentrations are both appropriately suppressed, but β -OHB concentrations are not, indicating that ketogenesis may not be impaired in the absence of growth hormone. The hypoglycaemia in this setting should be treated with both glucose and hydrocortisone.

Hypoglycaemia is uncommon in primary adrenal insufficiency (Addison disease). It can be precipitated by missing a meal and by exercise and is mainly due to impaired gluconeogenesis. Alcohol, by further impeding gluconeogenesis, may also provoke hypoglycaemia in these patients. A plasma cortisol concentration at the lower end of the reference range, during a spontaneous hypoglycaemic episode, is not sufficient evidence of adrenocortical insufficiency as it may be the result of a lower glycaemic threshold to stimulate cortisol secretion in patients with recurrent hypoglycaemia but without adrenal insufficiency. Plasma insulin and C-peptide concentrations are appropriately low and β -OHB concentrations are high. Hypoglycaemia occurring in acute adrenal insufficiency (Addisonian crisis) is a medical emergency and requires immediate correction by intravenous infusion of glucose in addition to hydrocortisone and other measures.

Congenital adrenal hyperplasia is a rare cause of neonatal hypoglycaemia. In the majority of patients, cortisol deficiency is not severe owing to a compensatory increase in ACTH secretion.

Although untreated hypothyroidism is associated with some lowering of fasting blood glucose concentrations, symptomatic hypoglycaemia has only been reported in a few cases and its existence has been questioned. Deficiency of other glucogenic hormones, such as adrenaline and glucagon, does not seem to cause hypoglycaemia.

Drug-induced hypoglycaemia

Insulin, insulin analogues or insulin secretagogues (sulfonylureas and meglitinides) are by far the most common causes of hypoglycaemia. Precipitating factors are restricted carbohydrate intake and liver and renal impairment. Sulfonylurea drugs can cross the placental barrier and stimulate insulin secretion in the fetus. Life-threatening hypoglycaemia has been reported in newborn infants of diabetic mothers who were treated with chlorpropamide during the third trimester. Sulfonylurea-induced hypoglycaemia may be prolonged, especially in those with renal impairment, lasting for up to seven days and necessitating continuous treatment. Mortality is high, ranging from 7.5 to 8.4%, and in attempted suicide with sulfonylurea drugs, the mortality is even higher with seven deaths out of 20 reported cases.

Salicylate overdose has been associated with hypoglycaemia in children. In adults, therapeutic doses of salicylates have been shown to lower blood glucose concentration in both diabetic and non-diabetic patients. The mechanisms by which salicylates produce hypoglycaemia are unknown. Enhancing insulin secretion and inhibiting hepatic gluconeogenesis are possible mechanisms.

Non-selective β -blockers (e.g. propranolol), in therapeutic doses, may induce hypoglycaemia, particularly in the presence of other precipitating factors such as liver disease, fasting or strenuous exercise. Hypoglycaemia has also been reported in newborn infants of mothers who were treated with propranolol until hours before delivery. The mechanism is thought to be through the prevention of the normal glucagon-mediated glycogenolytic and gluconeogenic responses by the liver. In diabetic patients, β -blockers may change the pattern of adrenergic symptoms of hypoglycaemia by diminishing the occurrence of tremor and palpitations and by increasing the occurrence of sweating. In addition, because glycogenolysis and gluconeogenesis in liver are stimulated through β_2 -receptors, blockade of these receptors by non-selective β -blockers could prolong recovery time from hypoglycaemia. Selective β_1 -blockers (e.g. atenolol and bisoprolol), however, do not impair the recovery from hypoglycaemia in patients treated with insulin or oral hypoglycaemic agents.

Quinine stimulates insulin secretion and its intravenous use in the treatment of falciparum malaria has been associated with profound hypoglycaemia. Plasma insulin and C-peptide concentrations are both elevated, indicating increased endogenous secretion. Falciparum malaria, especially in children, can itself cause hypoglycaemia. This, however, is associated with suppressed plasma insulin concentrations and is probably due to high glucose uptake by the parasitized erythrocytes. Acute kidney injury, hepatic dysfunction and starvation are common in malaria and may be additional factors in precipitating the hypoglycaemia. Symptoms of hypoglycaemia may be mistaken for those of cerebral malaria, and blood glucose should always be monitored in these patients. Life-threatening, quinine-induced hypoglycaemia can be treated with octreotide, which blocks the insulinotropic effect of quinine.

Pentamidine, a drug used for the treatment of trypanosomiasis and leishmaniasis, has a direct toxic effect on

β -Cells, causing release of preformed insulin and hence hypoglycaemia. This can be followed by β -Cell destruction with ultimate insulin deficiency and diabetes. Pentamidine is also used for the treatment of *Pneumocystis carinii* infection, and there are at least 30 case reports of severe pentamidine-induced hypoglycaemia in undernourished patients with AIDS (acquired immune deficiency syndrome).

Disopyramide, an antiarrhythmic drug, has also been associated with severe hypoglycaemia, particularly in elderly patients with renal or hepatic impairment. Patients with non-functional endocrine tumours and liver metastases may be at risk of developing hypoglycaemia when treated with long-acting somatostatin analogues, as a result of reduced glucagon and GH secretion.

Other drugs that have occasionally caused hypoglycaemia include angiotensin-converting-enzyme (ACE) inhibitors, angiotensin-receptor antagonists, haloperidol, lidocaine, *p*-aminobenzoic acid, phenylbutazone, propoxyphene, saquinavir, tricyclic antidepressants, tyrosine kinase inhibitors and sulphonamides.

Alcohol-induced hypoglycaemia

Alcohol can be associated with both fasting and reactive hypoglycaemia. Several mechanisms may be operational in alcohol-induced fasting hypoglycaemia but the most important is direct inhibition of gluconeogenesis. This is mainly due to accumulation of NADH and increased NADH/NAD⁺ ratio resulting from the oxidation of ethanol.

Alcohol-induced fasting hypoglycaemia characteristically occurs 6–36 h after ingestion of moderate to large amounts of alcohol. Most patients are aged between 20 and 40 years, but occasionally it can occur in children after a relatively small amount of alcohol. Patients present with neuroglycopenic symptoms including stupor and coma. The patient's breath may smell of alcohol and the symptoms may be mistaken for acute alcoholic intoxication. Therefore, delayed recovery from a presumed alcoholic intoxication should alert the doctor to the possibility of hypoglycaemia.

The blood glucose concentration is usually <2.2 mmol/L and alcohol is nearly always detectable in the blood, though the concentration is not necessarily very high. Severe metabolic acidosis with high blood lactate concentration is a characteristic feature. Hyperketonaemia and ketonuria are almost invariably present but predominantly in the form of β -OHB, since the accumulation of NADH suppresses the conversion of β -OHB to acetoacetate. Ketosis may, therefore, go unrecognized if methods that only detect acetoacetate are employed (e.g. Ketostix[®]). Plasma insulin and C-peptide concentrations during the hypoglycaemia are usually appropriately suppressed. Prompt diagnosis and treatment, with intravenous glucose, is essential since mortality is relatively high. Glucagon is not effective because hepatic glycogen stores are depleted by the time hypoglycaemia ensues.

Alcohol potentiates the hypoglycaemic effect of insulin and sulfonylurea drugs. In insulin-treated patients, ingestion of alcohol may produce profound hypoglycaemia, which can be fatal. Hypoglycaemia resulting from the combined effect of alcohol and sulfonylurea drugs tends

to be less profound, possibly because diabetic patients receiving these drugs tend to be obese and, therefore, are in part protected from the hypoglycaemic effects of alcohol. This is because alcohol does not inhibit the release of glucose from pre-existing glycogen stores, and in obese subjects, these stores are generally sufficient to meet the need for glucose during fasting for 12 h or more.

Alcohol has been shown to potentiate the insulin-stimulating effect of glucose and thus increases the risk for reactive hypoglycaemia. This reaction can be demonstrated in 10–20% of healthy subjects who have consumed a mixture of alcohol and sucrose, such as ‘gin and tonic’, on an empty stomach and who refrained from eating for a few hours afterward. This effect is not observed, however, when saccharin or fructose is substituted for sucrose as the sweetening agent. Starchy foods, such as nibbles and bread, increase the risk for reactive hypoglycaemia, whereas foods providing mainly fat or protein have the reverse effect.

Hypoglycaemia due to deficient energy intake

Symptomatic hypoglycaemia is well recognized in starvation and has been observed in patients with protein-calorie malnutrition and anorexia nervosa. The hypoglycaemia is usually due to reduced hepatic glycogen reserves as well as a reduced supply of gluconeogenic substrates. In addition, the reduction in ketosis due to markedly depleted fat stores deprives the central nervous system of an alternative source of energy. The major risk factors are low body weight and intercurrent infection. Compulsive excessive exercise in patients with anorexia nervosa may also be a factor. Plasma insulin and C-peptide concentrations are appropriately suppressed. Despite low circulating insulin concentrations, β -OHB concentrations are low, owing to depleted fat stores. Hypoglycaemia in anorexia nervosa has a poor prognosis.

Septicaemia

Bacterial septicaemia, especially Gram-negative, can occasionally cause hypoglycaemia. It has been postulated that cytokines produced by macrophages, in response to endotoxin stimulation, may induce hypoglycaemia by increasing insulin secretion. The endotoxins released may also have a direct hypoglycaemic effect, probably by inhibiting gluconeogenesis. Acute kidney injury, which is often associated with septicaemia, may also be an important factor in the pathogenesis of the hypoglycaemia.

Exercise-related hypoglycaemia

Exercise is associated with a marked increase in glucose uptake by muscles. During the first 5–10 min of severe exercise, glucose is supplied by the breakdown of muscle glycogen, but by 40 min, 75–90% of the glucose is supplied by the blood, mainly from increased hepatic glucose production. Initially, 75% of increased hepatic glucose output is derived from glycogenolysis and 25% from gluconeogenesis. With prolonged exercise, gluconeogenesis becomes more important and contributes up to 45%

of the total hepatic glucose output. Decreased plasma insulin and increased catecholamines, glucagon, cortisol and GH concentrations all contribute to the increased hepatic glucose output. If hepatic glucose production is inadequate, blood glucose cannot be maintained during exercise and hypoglycaemia ensues.

Hypoglycaemia following excessive exercise is well recognized and may be severe enough to cause seizures. Although exercising to exhaustion can produce hypoglycaemia, the symptoms of exhaustion are not related to hypoglycaemia and glucose administration does not modify the time of exercise to exhaustion.

Postprandial (reactive) hypoglycaemia

Rarely, hypoglycaemia occurs only in response to ingestion of a meal. Symptoms are usually neurogenic or autonomic and generally occur 2–4 h after ingestion of food and last for about 10–20 min. These symptoms are different from those of the dumping syndrome (flushing, sweating, abdominal cramps and hypotension), which occur within half an hour of eating. The diagnosis of postprandial hypoglycaemia requires the demonstration of the Whipple’s triad.

Postprandial hypoglycaemia may occur in patients who have undergone major gastric surgery, such as Roux-en-Y gastric bypass, and is also a feature of the autoimmune insulin syndrome, non-insulinoma pancreaticogenous hypoglycaemia, or may be alcohol induced or idiopathic (see below). Postprandial hypoglycaemia is also seen in patients with hereditary fructose intolerance after ingestion of fructose. It is worth noting that conditions that are usually associated with fasting hypoglycaemia, such as insulinoma, may occasionally produce symptoms only postprandially.

‘Idiopathic’ postprandial reactive hypoglycaemia has been documented in a small number of patients in whom no obvious other cause for hypoglycaemia was found. In these patients, a diet low in refined carbohydrates and high in soluble fibres is usually sufficient to relieve the symptoms. The addition of snacks in the middle of the morning and the afternoon may also prevent the fall in blood glucose. Patients should avoid rapidly absorbable sugars and should replace sucrose with either saccharin or fructose, a non-insulinotropic carbohydrate. If symptoms persist despite following the above dietary modifications, the use of an intestinal α -glucosidase inhibitor, such as acarbose, has been shown to be successful in alleviating symptoms in some patients with idiopathic postprandial reactive hypoglycaemia.

The postprandial syndrome

Patients who have postprandial symptoms but without documented hypoglycaemia should be referred to as having ‘the postprandial syndrome’. Symptoms are generally non-specific, such as feelings of vague ill health, lightheadedness, dizziness, anxiety, palpitations, fatigue, hunger and ‘inner trembling’, which may resemble the autonomic symptoms of hypoglycaemia. Many patients, especially those who are well informed about

hypoglycaemia from the lay literature, state that symptoms are relieved by eating, thereby suggesting hypoglycaemia. However, detailed questioning often reveals that symptoms are relieved either immediately or within a few minutes of eating foods such as cheese or bread, which are unlikely to raise blood glucose concentrations in such a short period of time. Many patients complain of generalized weakness and inability to concentrate in between attacks. The symptoms are not progressive but usually persist for many years. Reliance on the use of the oral glucose tolerance test (OGTT) in the investigation of this condition in the past has led to the misdiagnosis of 'reactive' hypoglycaemia in many of these patients. The OGTT should not be used in this context since 10% of healthy individuals may have low blood glucose concentrations (<2.6 mmol/L) during the test. In addition, in many patients with postprandial symptoms, the symptoms do not correlate with the nadir blood glucose concentrations and some individuals may even experience symptoms after a placebo oral glucose tolerance test. Self-monitoring of blood glucose at home, to document hypoglycaemia, using glucose meters should not be encouraged, since self-diagnosis by patients themselves is seldom confirmed by accurate investigations.

Inherited metabolic disease

Several inborn errors of metabolism can produce hypoglycaemia as one of their major clinical manifestations. These disorders are almost invariably diagnosed in childhood (see Chapter 24), but in mild cases, the diagnosis may not be made until middle life. The hypoglycaemia is usually of the fasting type, but in some disorders, for example galactosaemia and hereditary fructose intolerance, it only occurs following the ingestion of certain types of food.

CONCLUSION

Hypoglycaemia literally means low blood glucose concentration. Many apparently healthy subjects may have low blood glucose concentrations during a prolonged fast, or 3–5 h after ingestion of glucose. Pathological hypoglycaemia should, therefore, be defined as a clinical syndrome in which symptoms or signs of hypoglycaemia occur in the presence of low plasma glucose concentration and that the symptoms or signs are relieved after the plasma glucose concentration is raised

(Whipple's triad). Although it is frequently suspected as a cause of symptoms, hypoglycaemia is rare, other than in diabetic patients who are being treated with insulin or oral hypoglycaemic drugs.

The diagnosis of a hypoglycaemic disorder requires a high level of suspicion, a thorough history and a careful assessment of possible underlying illnesses or drugs. A healthy-appearing patient with hypoglycaemia requires a different diagnostic approach from a patient who has a concurrent illness or is hospitalized. Generally, the measurement of plasma glucose, insulin, C-peptide and proinsulin concentrations during symptoms may be sufficient to confirm the diagnosis and establish the cause of hypoglycaemia. In the majority of patients, however, further investigations (insulin autoantibodies, screening for hypoglycaemic agents) and provocation tests (prolonged fast or a mixed meal test) may be necessary. An inability to demonstrate a low plasma glucose concentration when the patient is symptomatic virtually excludes the diagnosis of hypoglycaemia.

Further reading

Cryer PE, Axelrod L, Grossman AB et al. Evaluation and management of adult hypoglycaemic disorders: an endocrine society clinical practice guideline. *J Clin Endocrinol Metab* 2009;94:709–28.

The most up-to-date evidence-based guidelines produced by the Endocrine Society and co-sponsored by the American Diabetes Association, European Association for the Study of Diabetes and European Society of Endocrinology.

Service GJ, Thompson G, Service FJ et al. Hyperinsulinemic hypoglycemia with nesidioblastosis after gastric bypass surgery. *N Engl J Med* 2005;353:249–54.

A report of six patients with postprandial neuroglycopenia after Roux-en-Y gastric bypass surgery.

Tesfaye N, Seaquist E. Neuroendocrine responses to hypoglycemia. *Ann N Y Acad Sci* 2010;1212:12–28.

A comprehensive review of the counterregulatory response to hypoglycaemia.

UK Hypoglycaemia Study Group. Examining hypoglycaemic risk in diabetes: effect of treatment and type of diabetes. *Diabetologia* 2007;50:1140–7.

An observational prospective study commissioned by the Department for Transport to compare rates of hypoglycaemia between patients with type 2 diabetes and type 1 diabetes in relation to treatment modalities and their duration.

Vezzosi D, Bennet A, Maiza JC et al. Diagnosis and treatment of insulinomas in the adults. In: Akin F, editor. *Basic and clinical endocrinology up-to-date*. Intech Open Access Publishing; 2011. <http://www.intechopen.com/books/basic-and-clinical-endocrinology-up-to-date/diagnosis-and-treatment-of-insulinomas-in-the-adults> [Accessed October 2013].

An excellent and detailed up-to-date review of the diagnosis, differential diagnosis and the medical management of adult patients with insulinoma.

Hypothalamic, pituitary and adrenal disorders

Miles J. Levy • Trevor A. Howlett

CHAPTER OUTLINE

INTRODUCTION 349

CLINICAL ANATOMY OF THE PITUITARY AND HYPOTHALAMUS 350

PHYSIOLOGY OF HYPOTHALAMO–PITUITARY–END ORGAN AXES 350

CLINICAL ANATOMY AND PHYSIOLOGY OF THE ADRENALS 352

ASSESSMENT OF NORMAL PITUITARY FUNCTION 352

Basal hormonal investigations 352

Dynamic tests of ACTH–adrenal function 354

Cortisol normal ranges, borderline responses, assay precision and dynamic test reproducibility 355

Assessment of growth hormone reserve 355

Releasing hormone tests 356

Other tests of gonadotrophin secretion 356

Dynamic tests of posterior pituitary function 357

Summary 357

A clinical approach to assessment of the whole ACTH–adrenal axis 358

Monitoring of pituitary function in disease states 359

OTHER DIAGNOSTIC TECHNIQUES IN PITUITARY DISEASE 360

Clinical assessment 360

Pituitary imaging techniques 360

PITUITARY HYPERSECRETION STATES 360

Pituitary adenomas 360

Prolactinoma 360

Acromegaly 362

Diagnosis and differential diagnosis of Cushing syndrome 363

Imaging 366

Outline of management 366

Thyroid stimulating hormone-secreting adenomas 366

Gonadotrophin-secreting adenomas 367

HYPOTHALAMIC AND PITUITARY DEFICIENCY STATES 367

Diseases that may lead to generalized hypopituitarism 367

Growth hormone deficiency 367

Gonadotrophin deficiency 368

Other isolated anterior pituitary deficiencies 369

Diabetes insipidus 369

ADRENAL DISEASE 369

Clinical features of Addison disease 369

Congenital adrenal hyperplasia 370

Assessment of adrenal incidentaloma 370

MONITORING PITUITARY AND ADRENAL REPLACEMENT THERAPY 370

CONCLUSION 371

APPENDIX 371

INTRODUCTION

The investigation of disorders of the pituitary and hypothalamus and of adrenocortical function frequently causes great apprehension among non-endocrinologists, whether they are clinical biochemists or general physicians, because of their supposed complexity and the frequency of atypical, borderline or artefactual results. It is certainly true that the hypothalamo–pituitary axis plays the central role in the control of a large number of hormonal systems and of many other important aspects of homeostasis. However, the clinically relevant

physiology of each hormonal axis is well understood and the plasma concentrations of the majority of the hormones secreted by both the pituitary and the relevant peripheral endocrine glands can now be readily and accurately measured in most clinical biochemistry laboratories. Thus, a logical approach to the measurement of the relevant plasma hormone concentrations, applying basic physiological principles, can result in a relatively simple and efficient assessment of pituitary function in the majority of cases.

This chapter will outline a clinically relevant and efficient approach to the assessment of pituitary and adrenal

function. In many cases, simple measurement of basal, unstimulated plasma concentrations of the appropriate hormone(s) will give all the information required for clinical management. Some of the traditionally used dynamic endocrine tests can now be reasonably discarded, while others remain essential, and the choice of these will be discussed critically. Recommendations for clinical management of borderline or contradictory results will also be discussed.

CLINICAL ANATOMY OF THE PITUITARY AND HYPOTHALAMUS

The pituitary gland (Fig. 18.1) lies within a bony compartment, the pituitary fossa or sella turcica, beneath the hypothalamus to which it is connected by the pituitary stalk. The fossa is separated from the subarachnoid space and cerebrospinal fluid by the diaphragma sellae. The adult human pituitary consists of two lobes (the third, neurointermediate lobe of other species is not present except in fetal life). The anterior lobe (adenohypophysis) is embryologically derived from cells of Rathke's pouch. It contains a variety of cell types differentiated to secrete the majority of the peptide and glycopeptide hormones that control the function of peripheral endocrine organs: corticotrophs (adrenocorticotrophic hormone, ACTH), lactotrophs (prolactin, PRL), gonadotrophs (luteinizing hormone and follicle stimulating hormone: LH and FSH), thyrotrophs (thyroid stimulating hormone, TSH) and somatotrophs (somatotrophin or growth hormone, GH). The anterior lobe has no direct arterial blood supply but is supplied by the portal vessels that arise in the median eminence of the hypothalamus. The secretion of anterior pituitary hormones is controlled by releasing and inhibiting factors, which are predominantly peptides, released into the portal circulation from the nerve terminals of a variety of hypothalamic neurons. These neurons in turn are subject to the effects of the modulatory neurotransmitters of other neurons, within the hypothalamus and beyond.

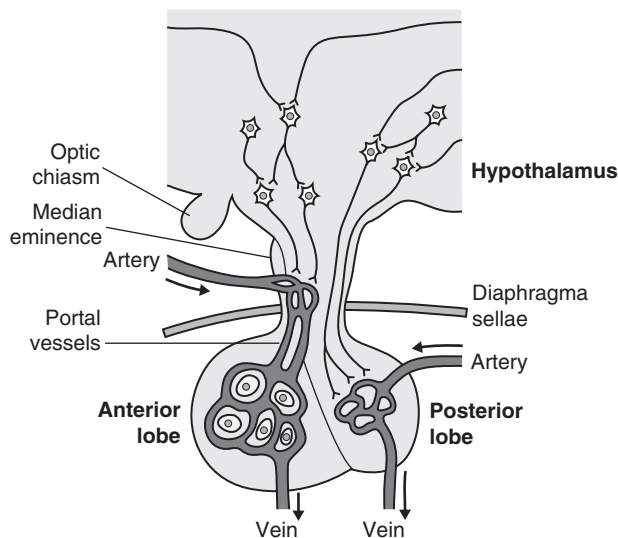


FIGURE 18.1 ■ Functional anatomy of the human pituitary and hypothalamus.

The critical role of the portal circulation in the control of anterior pituitary function means that any disease process that interferes with this blood supply, for example non-functioning tumours of the pituitary or hypothalamus, will often result in severe pituitary dysfunction, even if the actual hormone-secreting cells of the anterior pituitary have not themselves been destroyed.

The posterior lobe (neurohypophysis), in contrast, is embryologically derived from the brain, has a direct arterial blood supply and is not controlled via a portal circulation. Hormonal secretion occurs directly from the nerve terminals of vasopressin (antidiuretic hormone, ADH) and oxytocin neurons, whose cell bodies lie within the hypothalamus, primarily in the supraoptic and paraventricular nuclei.

Both the anterior and posterior lobes contain a variety of other cell types. These are not directly responsible for hormonal secretion into the peripheral circulation, although they may have important modulatory or paracrine roles in the control of pituitary function.

Many other important neural structures lie within or adjacent to the hypothalamus, notably the optic chiasm and the neural centres controlling thirst, osmoregulation, appetite, temperature homeostasis and circadian rhythm. Diseases of the pituitary and hypothalamus can result in dysfunction of any of these structures, in addition to the hormonal syndromes described here. In particular, visual field loss due to compression of the optic chiasm is a frequent finding with large pituitary tumours.

These anatomical relationships indicate that the pituitary gland is truly at the interface of the mind and body, with a central role in homeostasis, explaining its source of fascination for physiologist and physician alike.

PHYSIOLOGY OF HYPOTHALAMO–PITUITARY–END ORGAN AXES

The basic physiology of the important endocrine axes is summarized diagrammatically in Figures 18.2 and 18.3. Many additional releasing, inhibiting and modulating factors have been identified, primarily in animal experiments, but none of these has yet been shown to be of relevance clinically, nor for the understanding of the principles of pituitary function testing. Hormones that can be measured in the laboratory are indicated in the figures. Peripheral plasma concentrations of hypothalamic releasing factors can also be measured (with difficulty), but for the most part do not reflect hypothalamic activity and are not of clinical relevance.

Adrenocorticotrophic hormone (ACTH) stimulates adrenal cortisol secretion and is itself controlled by hypothalamic corticotrophin releasing factor (CRF). Corticotrophin releasing factor is now known to be a complex of factors including a 41-residue peptide (CRF-41 or corticotrophin releasing hormone, CRH) and vasopressin, which act synergistically. Adrenocorticotrophic hormone, and therefore cortisol, are secreted with a pronounced circadian rhythm: concentrations are low, typically undetectable, at midnight (if asleep), rise in the final hours of sleep to reach a peak shortly after wakening in the morning and steadily decline throughout the

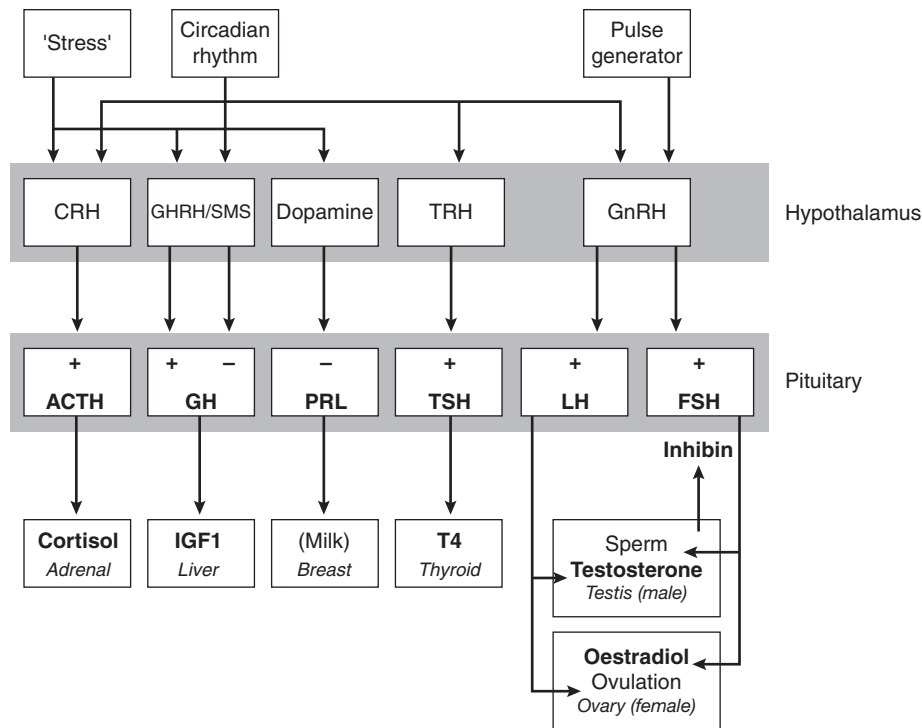


FIGURE 18.2 ■ Physiological relationships of the anterior pituitary. Hormones (in bold) can be readily measured in most clinical biochemistry laboratories. See text for explanation of abbreviations.

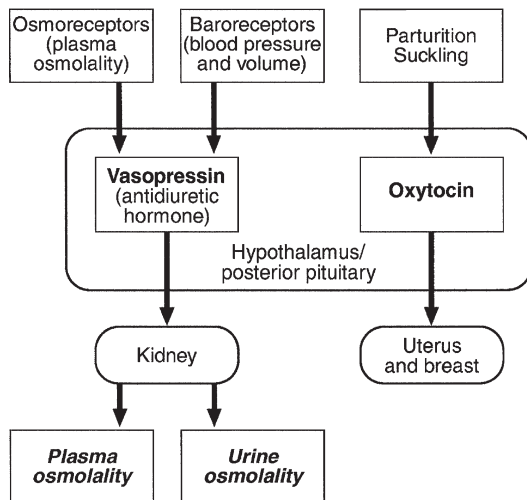


FIGURE 18.3 ■ Physiological relationships of the posterior pituitary. Osmolalities (italicized) can be readily measured in most clinical biochemistry laboratories. Hormones (emboldened) can be measured with difficulty in a few centres.

rest of the day with superimposed peaks of secretion with meals, exercise and stressful events. 'Stress' ('fight and flight', illness, hypoglycaemia, surgery etc.) is a major determinant of ACTH and cortisol secretion, which therefore means that basal concentrations in a well individual may not always reflect the ability to respond to illness, and conversely that high concentrations in an individual with other significant disease do not always imply oversecretion.

Secretion of growth hormone (GH) is controlled by complementary stimulatory and inhibitory factors, GH

releasing hormone (GHRH, a 44-residue peptide) and somatostatin (SMS, a cyclic, 14-residue peptide), respectively. Growth hormone secretion occurs in pulses, predominantly at night during the early hours of sleep. Only infrequent, small pulses occur during the day when plasma concentrations are usually undetectable, although GH secretion also occurs in response to stress. Growth hormone's effects on growth are mediated via synthesis of insulin-like growth factor 1 (IGF-1 – previously called somatomedin C), which is synthesized mostly in the liver but also in peripheral tissues.

Prolactin secretion is unique in being predominantly controlled by an inhibitory hypothalamic factor, dopamine. This fact explains why PRL deficiency is rare and why any disease that interferes with pituitary anatomy, and thus the flow of portal blood, may be associated with hyperprolactinaemia. Plasma PRL is raised physiologically during pregnancy and lactation, and release also occurs in response to stress. The physiological role of PRL in non-lactating women and in men is unknown.

Hypothalamic thyrotrophin releasing hormone (TRH), a tripeptide, controls TSH secretion, which in turn stimulates the synthesis and secretion of thyroid hormones thyroxine (T4) and tri-iodothyronine (T3). Secretion is regulated primarily to maintain normal circulating concentrations of thyroid hormones, although there is also a slight circadian rhythm.

Luteinizing hormone and FSH are both controlled by a single stimulatory decapeptide, gonadotrophin releasing hormone (GnRH or luteinizing hormone releasing hormone, LHRH). The major role of LH in the male is to stimulate testosterone secretion by testicular Leydig cells, and in the female to stimulate ovulation at mid-cycle. Follicle stimulating hormone controls spermatogenesis in the male and ovarian follicular development,

and therefore oestradiol secretion, in the female, and in both sexes stimulates gonadal secretion of inhibin, which is responsible for some negative feedback control. The most important feature of the secretion of GnRH, and thus LH and FSH, is its pulsatile nature, imposed by the hypothalamic pulse generator. It is likely that differential secretion of LH and FSH is achieved by modulation of the frequency and amplitude of pulsatile GnRH secretion, and the peripheral effects of LH and FSH are critically dependent on appropriate pulsatile variation in concentrations. Gonadotrophin releasing hormone pulses occur, on average, every 90 min, but in the female are more rapid in the follicular phase and slower in the luteal phase of the menstrual cycle. During early puberty, the majority of pulses occur at night, during sleep.

Vasopressin (antidiuretic hormone, ADH) secretion is controlled primarily by osmotic and blood volume sensing mechanisms (osmoreceptors and baroreceptors, respectively). High plasma osmolality and hypovolaemia or hypotension are the major stimulatory factors. Vasopressin acts directly on the kidney to increase water reabsorption in the renal collecting ducts and thus reduce urine volume. Corticosteroids are necessary for the excretion of free water, so deficiency may mask the symptoms of ADH deficiency.

Oxytocin is released during labour and during suckling, but mechanisms controlling its secretion have not been studied extensively in humans and there appears to be no clinical syndrome related to oxytocin deficiency.

Secretion of all pituitary hormones is controlled, to a greater or lesser extent, by negative feedback. This occurs at all levels, including the direct inhibition by a hormone of its own secretion ('ultra-short-loop'), but the most important negative effects are generally those of the peripheral 'end hormone' on hypothalamic and/or pituitary secretion. The existence of such negative feedback is of crucial importance to the interpretation of basal hormone concentrations in pituitary disease and is the basis of a number of dynamic tests of pituitary function.

CLINICAL ANATOMY AND PHYSIOLOGY OF THE ADRENALS

The two adrenal glands lie superior to the kidneys and are composed of the inner catecholamine-secreting medulla, which is controlled centrally by direct innervation, and the outer cortex, which is controlled by classic endocrine pathways. Glucocorticoid and androgen secretion is predominantly from the zona fasciculata and zona reticularis and controlled by pituitary ACTH secretion, while mineralocorticoid secretion is predominantly from the zona glomerulosa and controlled via the renin-angiotensin system (see Chapter 4). This means that the patterns of adrenal deficiency vary with aetiology: thus a disease that destroys the adrenal itself (e.g. autoimmune Addison disease, adrenal tuberculosis or bilateral adrenalectomy) results in deficiency of both glucocorticoid and mineralocorticoid secretion, whereas ACTH deficiency from pituitary disease results in glucocorticoid deficiency alone and, thus, less marked disturbances of salt and water balance. Even complete adrenalectomy does not seem to cause a clinical

deficiency of catecholamines, since secretion by the rest of the sympathetic nervous system appears to compensate for the loss of adrenal secretion.

Adrenal steroid synthesis involves a complex and intersecting sequence of enzymatic steps within the cytoplasm and mitochondria of the adrenal cell. Pathways are illustrated in Figure 18.4. The initial, rate-limiting, step in steroidogenesis in the normal subject is the conversion of cholesterol to pregnenolone. Adrenocorticotrophic hormone has rapid effects on cholesterol transport into mitochondria and longer-term effects on transcription of genes encoding a number of enzymes in the synthetic pathway. These pathways are of clinical significance in congenital adrenal hyperplasia, where inherited defects in a variety of these enzymes give rise to the various subtypes and clinical syndromes (see Table 18.1), and in clinical pharmacology, since metyrapone, a drug that inhibits 11 β -hydroxylase, can be used to decrease cortisol synthesis in Cushing syndrome.

ASSESSMENT OF NORMAL PITUITARY FUNCTION

Basal hormonal investigations

A single, basal blood sample, especially if taken at 09.00 h, can give extensive and, for some purposes, complete information regarding pituitary function. The blood should ideally be taken under resting, unstressed conditions but, while the stress of major illness or surgery significantly alters results and thus might cause diagnostic errors when investigating pituitary hyperfunction, minor anxiety regarding hospital attendance or venepuncture probably has little effect.

The pituitary-thyroid axis is fully assessed by the simultaneous measurement of plasma free thyroid hormone concentrations and TSH. Thyroid stimulating hormone deficiency is characterized by a low plasma free T4 concentration without the expected elevation of plasma TSH. In hypopituitarism, TSH may be low (although rarely undetectable) or within the normal range (inappropriately for the low T4), which represents the most potent argument against the use of TSH as the sole screening test for thyroid function in a laboratory.

Basal plasma PRL gives full physiological information regarding the function of its axis. Concentrations frequently vary from day to day (particularly minor elevations), so that 2–3 basal samples are usually necessary before any treatment is instituted.

In the male, a normal basal plasma testosterone (preferably taken at 09.00 h, in view of the significant circadian rhythm) effectively demonstrates normal LH secretion, and demonstration of a normal sperm count would confirm normal FSH secretion (although this is rarely performed unless infertility is an issue). Gonadotrophin deficiency is diagnosed by the combination of a low plasma testosterone with low or inappropriately 'normal' (rather than elevated) concentrations of LH and FSH. The significance of borderline testosterone concentrations – near or just below

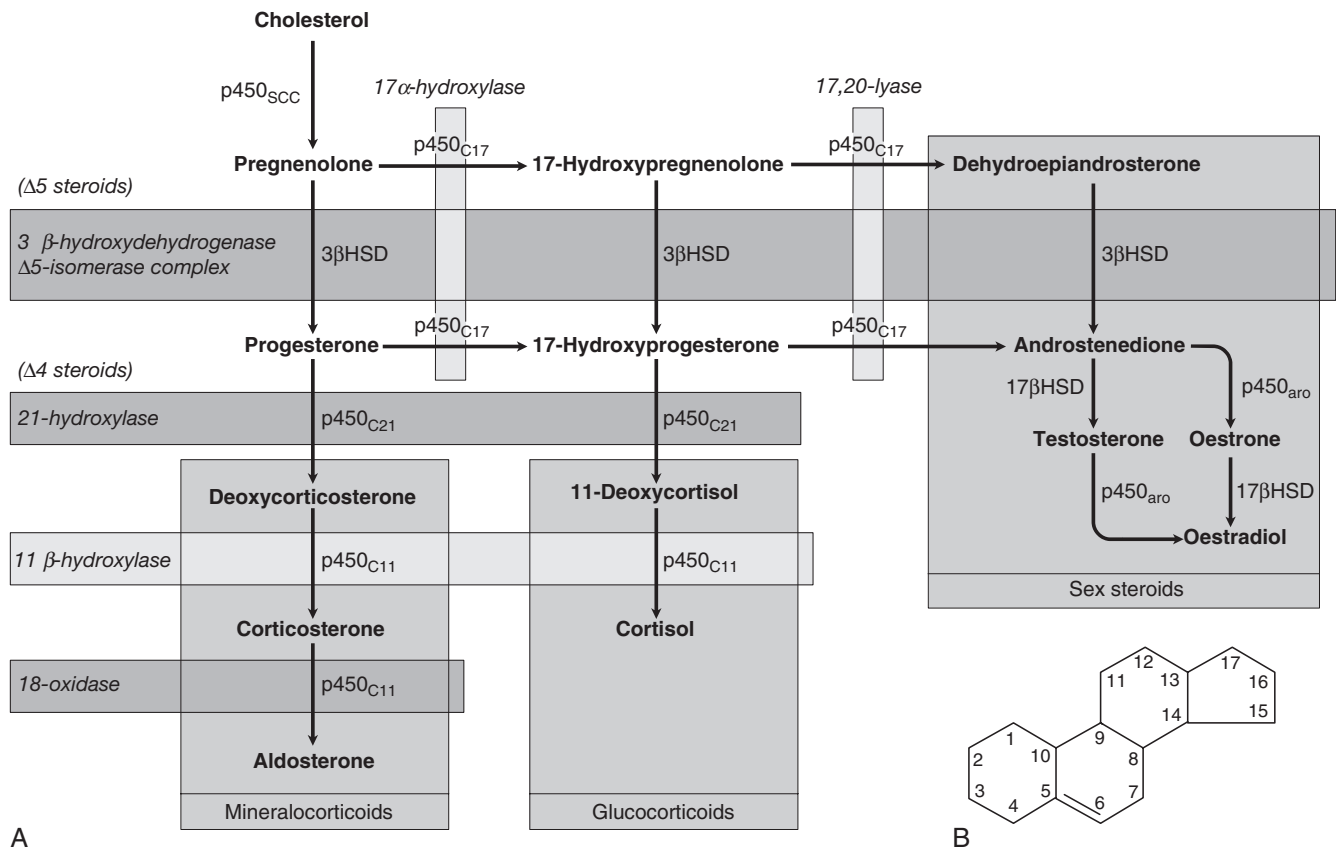


FIGURE 18.4 ■ (A) The pathways of steroid hormone synthesis within the adrenals (and gonads). Enzymes are named by reaction catalysed (e.g. 21-hydroxylase) or by enzyme name (e.g. p450_{C11}). p450 enzymes are in mitochondria and other enzymes are in the cytoplasm; steroid synthesis involves repeated passage of precursors between these cellular compartments. (B) Basic structure of the steroid molecule (from Kumar and Clark 2012 Clinical medicine. 8th ed. Edinburgh: Churchill Livingstone, with permission).

TABLE 18.1 Subtypes of congenital adrenal hyperplasia

Enzyme defect	Frequency	Clinical features
21-hydroxylase p450 _{C21} /CYP21	95% of cases 1 in 15000 live births	Ambiguous genitalia in females Virilization, some salt losing
11-β-hydroxylase p450 _{C11} /CYP11B	<5% of cases	Ambiguous genitalia in females Virilization, hypertension
17-α-hydroxylase p450 _{C17}	Rare	Ambiguous genitalia and lack of virilization in males Hypertension
3β-hydroxysteroid dehydrogenase 3βHSD/HSD3B2	Rare	Ambiguous genitalia in males Mild virilization in females Occasional Addisonian crises
Cholesterol desmolase p450 _{SCC} /CYP11A	Rare	Addisonian crises

the lower end of the reference range – together with normal LH and FSH, particularly in the ageing male, is an area of growing clinical interest but still with considerable controversy and is discussed more fully later in this chapter.

Assessment of the pituitary–gonadal axis in the female is complicated by the physiological changes during the menstrual cycle. A normal luteal phase plasma progesterone is the ultimate biochemical test of the adequacy of the entire axis. Spontaneous, normal, regular menstruation also excludes gonadotrophin deficiency without the need for additional biochemical tests. Gonadotrophin deficiency is suggested in the context of amenorrhoea by the presence of a consistently low plasma LH and FSH, although a variety of ‘hypothalamic’ causes of amenorrhoea also show this combination (see below). In postmenopausal women (or, in the context of longstanding amenorrhoea, arbitrarily women aged >50 years), who are not on oestrogen replacement, the simple absence of the usual menopausal elevation of LH and FSH is sufficient to diagnose gonadotrophin deficiency.

Basal plasma cortisol is not always sufficient to diagnose ACTH deficiency or normality, because of the circadian rhythm and since concentrations that need to be achieved during physiological stress are much higher than those found under normal basal circumstances.

Nevertheless, a basal cortisol >550 nmol/L in an unstressed individual certainly excludes ACTH deficiency and, in our experience, a 09.00h cortisol <100 nmol/L is never, and >400 nmol/L is always, associated with a normal ACTH reserve (see below).

Basal concentrations of GH are usually unhelpful in the diagnosis of deficiency, since basal concentrations in normal individuals are usually undetectable. A low serum IGF-1 (using age-related normal ranges) is a good predictor of severe GH deficiency but, conversely, an IGF-1 in the normal range does not exclude severe deficiency that might respond to replacement therapy.

Plasma vasopressin and oxytocin concentrations are measured by only a few centres and are not used routinely in the assessment of basal pituitary function. Measurement of paired plasma and urine osmolality on a basal early morning sample does, however, give valuable information. A urine osmolality >600 mmol/kg excludes diabetes insipidus (DI) as long as plasma osmolality is normal (280–295 mmol/kg). Conversely, DI is suggested by an elevated plasma osmolality (>300 mmol/kg) with a urine osmolality <600 mmol/kg. Other combinations are non-diagnostic and require a water deprivation test if DI is suspected clinically. Oxytocin secretion is never investigated clinically.

The order of progression of hypopituitarism with progressively severe pituitary disease is usually predictable. Growth hormone and gonadotrophin secretion are usually most susceptible to damage, followed by ACTH and TSH, with PRL deficiency occurring only rarely and usually only after extensive pituitary damage by disease or surgery. It is thus unusual, although not impossible, for a patient with regular menstruation or normal male gonadal function to develop ACTH deficiency first in the course of their disease. Diabetes insipidus usually only occurs with inflammatory or invasive pituitary disease, after surgery or with diseases primarily involving the hypothalamus. Thus clinical background may influence the degree of investigation considered appropriate in an individual patient with borderline or equivocal results in a single axis.

Dynamic tests of ACTH–adrenal function

Insulin stress test

The insulin stress test (IST) (or insulin tolerance test, ITT) has traditionally been the ‘gold standard’ test to assess the adequacy of the ACTH–adrenal axis. After adequate hypoglycaemia (blood glucose <2.2 mmol/L), a normal rise of plasma cortisol (traditionally to >550 nmol/L) excludes ACTH and cortisol deficiency. Patients with such a response are capable of producing a normal adrenocortical response to stressful illness, without the need for replacement therapy. The increment of plasma cortisol does not add additional information. Patients with borderline peak responses (400–550 nmol/L) may only require steroid replacement during stressful illnesses, but patients with lower responses will typically be symptomatic and almost always require standard replacement therapy.

An IST is contraindicated by cardiac disease (an ECG should be normal) and by epilepsy (or other unexplained

blackouts) and always requires close medical and nursing supervision to maintain safety. For this reason, alternative, simpler tests have been sought, although none has yet achieved universal acceptance. A retrospective audit of 230 ISTs in our own laboratory indicated that patients with a basal cortisol <100 nmol/L never, and >400 nmol/L always, demonstrated a normal cortisol response, so that an IST is probably unnecessary in such patients (representing 50% of ISTs performed in our study) if the only aim is to assess pituitary–adrenal reserve. In our experience, patients who have recently discontinued steroid replacement with a basal cortisol <200 nmol/L also never have a normal response.

Short tetracosactide (synacthen, tetracosactrin, ACTH) test

This test, involving administration of tetracosactide (synthetic ACTH_{1–24}), is the cornerstone for diagnosis of primary adrenal failure, but is increasingly used in the context of pituitary disease. It has long been accepted that a normal cortisol response to tetracosactide excludes Addison disease, but to test for ACTH deficiency by administration of ACTH is clearly illogical on purely physiological grounds, since it directly tests only the adrenal response and not that of the pituitary or higher centres. However, presumably since ACTH deficiency leads to adrenal atrophy, a number of workers have shown close correlations between the cortisol responses to a simple intravenous tetracosactide test and to an IST, and have suggested that the former is a satisfactory test of the ACTH–adrenal axis as long as the test is not done within four weeks of the pituitary insult.

A majority of UK endocrinologists now accept that a normal plasma cortisol response, 30 min after tetracosactide (250 µg), adequately excludes ACTH deficiency, and several long-term clinical observational studies have confirmed the safety of such an approach. This test was originally validated with i.v. tetracosactide but responses to i.m. injection are equivalent. However, normal responses are different (typically 150 nmol/L or so higher) at 60 min, so an appropriate normal range is needed for each time point and, in practice, we recommend only measuring the 30 min value to avoid confusion.

Since the standard 250 µg dose of tetracosactide is substantially supraphysiological, many workers have advocated use of a lower dose stimulation test (typically 1 µg). However, this approach is more complicated since no suitable dose preparation exists, and it has not been demonstrated to be significantly superior in clinical practice to the traditional test.

Other tests

The glucagon test (1 mg i.m.) has been advocated to assess ACTH and GH reserve when an IST is contraindicated (usually with the same criteria for a normal response). Glucagon frequently causes nausea and vomiting, normal individuals sometimes fail to show a ‘normal’ response and sampling must continue over a 4 h period, making this test relatively unattractive as a routine alternative.

The metyrapone test (750mg orally, four-hourly for 24h) has been widely used as a test of ACTH reserve, particularly in North America; it measures the response of urinary cortisol precursors or, more recently, plasma 11-deoxycortisol, to blockade of adrenal cortisol synthesis by this drug. The test is time consuming, often unpleasant for the patient and does not test the response to physiological stress.

An intravenous short tetracosactide test is now recognized by most endocrinologists to be a more appropriate and simpler alternative to either of these tests, unless information on GH reserve is also required.

Cortisol normal ranges, borderline responses, assay precision and dynamic test reproducibility

Normal ranges for cortisol, particularly for dynamic tests, were predominantly developed and validated using old assay technologies (e.g. fluorimetry), which are no longer in use, and although a variety of adjustments (typically lowering of cortisol cut-off concentrations for a normal response) have been suggested, these have rarely been completely validated.

Detailed studies of normal responses to tetracosactide have shown substantial differences with gender and between commonly used cortisol assays (see Clark et al. in Further reading, below). Most workers and most assays report a lower limit of the normal response between 500 and 600nmol/L, but some other published studies report lower peak responses than this in some apparently healthy individuals (lowest values ~400nmol/L). Similarly, some recent studies reporting cortisol responses to IST in healthy volunteers have reported peak responses in some individuals on some occasions between 400 and 500nmol/L.

Studies of test–re-test variability for both the tetracosactide and IST tests have mostly reported a coefficient of variation (CV) of around 10%. External quality control schemes also frequently report an interassay/intermethod CV for cortisol of 10% or higher. Therefore, if we assume conservatively that the combined variability from these two sources is represented by an overall CV of 10%, then simple statistics show that the 95% confidence limits for a test response with a ‘true’ value of 500nmol/L actually extends from 400 to 600nmol/L. These considerations potentially apply to both the tetracosactide test and the IST.

All of these factors lead the authors to define basal cortisol and cortisol responses to tetracosactide and other dynamic tests broadly rather than using a single precise cut-off, and to recognize a group with borderline or equivocal responses who cannot be clearly classified as ‘normal’ or ‘abnormal’:

- 100 nmol/L: a basal cortisol below this value almost always represents severe hypoadrenalism (or treatment with steroids)
- 400 nmol/L: a basal cortisol above this value indicates a normal hypothalamo–pituitary–adrenal (HPA) axis – unless the patient is acutely unwell or has ongoing unexplained symptoms consistent with

hypoadrenalism. A cortisol response to tetracosactide or IST below this concentration usually means that routine steroid replacement is required

- 600nmol/L: a cortisol response to tetracosactide to above this concentration at 30min almost always means that ACTH deficiency or primary adrenal failure is excluded. A cortisol response to above this concentration in an IST certainly excludes the need for steroid replacement even during severe intercurrent illness
- 100–400nmol/L: basal cortisol concentrations in this range need a dynamic test to confirm the abnormality of the axis; a tetracosactide or IST peak response in this range usually indicates a requirement for routine steroid replacement
- 400–600nmol/L: peak responses to tetracosactide or IST in this range in patients with known pituitary disease may not require the patient to receive regular replacement, but do require education of the patient about the possible need for replacement and symptoms of hypoadrenalism during intercurrent illness (and often provision of an emergency supply of hydrocortisone). Responses to tetracosactide in this range can sometimes be seen in normal individuals, so if there is no other evidence of adrenal or pituitary disease, further investigation may be required (e.g. renin and ACTH concentrations, possibly followed by IST to exclude isolated ACTH deficiency if the former are normal). Interpretation of isolated responses in this range are further complicated by the fact that patients with chronic fatigue syndrome may show borderline reduced cortisol dynamic responses, although in most cases they do not respond symptomatically to steroid replacement.

Assessment of growth hormone reserve

There are several biochemical tests that are used in clinical practice to confirm growth hormone deficiency (GHD). The clinical features of GHD are non-specific, and the different GH stimulation tests have different degrees of ability to stimulate GH release. For this reason, there continues to be debate about the most appropriate test and what extent of GH response constitutes true deficiency.

Insulin stress test

The insulin stress test (IST) remains the gold standard test of GH deficiency. In adults, a peak GH response to adequate hypoglycaemia (serum glucose <2.2mmol/L) of >7µg/L indicates a normal GH reserve, whilst a peak GH <3µg/L, as defined by the UK National Institute for Health and Care Excellence (NICE), indicates severe deficiency. In children, a peak GH <6.7µg/L after two provocative tests supports a diagnosis of growth hormone deficiency, although only one dynamic test is required in children with known hypothalamo–pituitary disease. The IST is unpleasant for patients and requires close medical supervision as it is potentially hazardous and is contraindicated in ischaemic heart disease and epilepsy. As well as

being a reliable GH stimulus, the IST is also a good assessment of ACTH reserve so it a useful way of assessing GH and ACTH reserve simultaneously.

Children of both sexes may show subnormal GH responses to hypoglycaemia (and other dynamic tests) in the years immediately preceding puberty (objectively when the bone age is >10 years). At the present time, there is no consensus with respect to the practice of 'priming' with gonadal steroids prior to such tests.

Other pharmacological tests

The glucagon test can be useful in the assessment of GHD when an IST is contraindicated, but it is a less reliable stimulus of GH and a poorer test of ACTH reserve than the IST, and its mechanism of action is not well understood. The criteria for diagnosing GHD in the glucagon test are the same as for the IST described above.

Other provocative tests of GH reserve include the GH-releasing hormone (GHRH)+arginine and GHRH+GH-releasing hexapeptide (GHRP-6) test, both of which have now been evaluated in large numbers of patients. The GHRH+arginine test may be particularly useful for young patients with radiation-induced GHD as occasionally the IST can give false negative results. Because GHRH+arginine is a more powerful GH stimulus, the cut-off point for diagnosing GHD is <9 µg/L.

In an attempt to avoid an IST, a variety of other pharmacological tests of GH reserve have also been used, particularly in children, including clonidine (0.15 mg/m²) and arginine infusion (0.5 g/kg). Use of these tests tends to be a matter of local preference; all are associated with rather variable responses and the peak GH values used to diagnose deficiency vary from laboratory to laboratory.

Exercise testing

Growth hormone is released during exercise and some laboratories have used an exercise test to assess GH reserve. Intensive, well controlled exercise (bicycle ergometer or treadmill) may exclude GH deficiency (by a peak GH response >6.7 µg/L), but less well controlled exercise ('run up and down the stairs') is unreliable.

Assessment of physiological growth hormone secretion

Because GH is secreted in a pulsatile fashion with increased pulses at night, serial blood tests, which can be done overnight or over 24h, can be used to monitor the frequency and amplitude of GH pulses. Because this approach is labour intensive, requiring hospital admission with multiple blood samples, it is mainly a research tool.

Measurement of urinary GH is not part of routine clinical practice because of its low concentration in urine, but sensitive assays may make this possible in the future, particularly for the surveillance of GH replacement therapy.

Insulin-like growth factor 1 is produced by the liver in response to GH and, although measurement of its concentration is a useful diagnostic adjunct to provocative

GH stimulation tests, it should not be used alone to diagnose GHD. However, the finding of a low IGF-1 in patients with three or four other pituitary deficiencies makes the likelihood of GHD sufficiently high that a formal provocative test may not be necessary.

Re-evaluation of GH status in young adults

In patients with isolated GHD that developed in childhood, it can be difficult to know whether GH replacement should be continued into adulthood. Patients with genetic hypopituitarism or multiple hormone deficiencies are highly likely to need to continue GH replacement. However, it may be able to be discontinued in patients with isolated or less severe GHD. In this situation, patients are re-tested during the period of transition between childhood and adulthood and the current consensus guidelines suggest a cut-off of <6 µg/L during an IST to indicate a continuing need for GH replacement. In this situation, the presence of clinical symptoms can also help the decision about whether to continue GH, which is best done by evaluation of the Adult Growth Hormone Deficiency Assessment (AGHDA) score. Consensus guidelines suggest that young adults with GHD should continue GH until the age of 25 years, when peak bone mass is likely to have been achieved.

Releasing hormone tests

Thyrotrophin releasing hormone (200 µg i.v.) and LHRH (100 µg i.v.) tests have been used traditionally in combination with the IST to assess pituitary reserve in the combined pituitary function test. Growth hormone releasing hormone testing (100 µg i.v.) and CRH testing (100 µg i.v.) have also more recently been advocated as additional tests.

All these releasing hormone tests assess only the readily releasable pituitary pool of their relevant hormones and do not assess the ability of the axis to respond to appropriate physiological stimuli. When the cause of hypopituitarism is at the hypothalamic level (which is frequently the case), then 'normal' TRH, LHRH, GHRH or CRH responses may be seen in the presence of deficiency documented by other means. Conversely, 'subnormal' responses are often seen when the axis is otherwise felt to be normal. Furthermore, a 'delayed' TRH test response (peak TSH at 60 min rather than 20 min), initially considered to be characteristic of hypothalamic disease, is not infrequently seen in primary pituitary disease.

For these reasons, the authors believe that all these releasing hormone tests should now play no part in routine pituitary function testing, but be reserved (if used at all) for the rare cases requiring differentiation between pituitary and hypothalamic deficiency states.

Other tests of gonadotrophin secretion

Clomifene test

When the differentiation of gonadotrophin deficiency from other causes of 'hypothalamic amenorrhoea' is in doubt, a clomifene test (3 mg/kg, maximum 200 mg, daily

for seven days) may give additional information. Patients with organic gonadotrophin deficiency show no increase in LH and FSH (measured on days 0, 4, 7 and 10), whereas some patients with hypothalamic amenorrhoea, weight-related amenorrhoea or delayed puberty show a gonadotrophin rise, usually followed by menstruation (about 28 days later). In early puberty, LH and FSH may show a characteristic fall rather than a rise.

Assessment of luteinizing hormone pulsatility

Measurement of plasma LH every 10–15 min for 6–8 h or more, when analysed by an appropriate computer algorithm, allows detailed assessment of LH pulse frequency and amplitude. This may give considerable insight into pituitary physiology but, in view of the time and number of assays involved, is essentially a research tool.

Dynamic tests of posterior pituitary function

Water deprivation test

Water deprivation is the standard physiological test of vasopressin secretion. The patient is observed, without access to water, with serial measurements of plasma and urine osmolality, urine volumes and body weight over a period of up to 8 h. If only borderline deficiency is suspected then it is wise, in order to shorten the procedure, to ask the patient to take no fluids from midnight before a morning test; otherwise the patient should be allowed free fluids prior to the test to avoid severe dehydration. Simple measurement of 24 h urine volume prior to the test may guide which approach is most appropriate; if this is $<2\text{ L}/24\text{ h}$ then DI is unlikely in the first place. If body weight falls by more than 3% during the test, this suggests severe dehydration. The test should be stopped, and plasma osmolality measured for confirmation before allowing access to fluids.

Normal vasopressin secretion is demonstrated by a urine:plasma osmolality ratio of $>2:1$ (or, more simply, a urine osmolality $>600\text{ mmol/kg}$), in the presence of a normal plasma osmolality (280–295 mmol/kg). A high plasma osmolality ($>300\text{ mmol/kg}$) with a urine osmolality of $<600\text{ mmol/kg}$ is diagnostic of DI. Intermediate responses are non-diagnostic and water deprivation must be continued or repeated, until diagnostic values are obtained. Once a diagnostic response has been obtained, then desmopressin (2 μg i.m.) is administered, free fluids are allowed and urinary osmolality is measured hourly for 2 h. In the presence of DI, full concentration of the urine after desmopressin indicates vasopressin deficiency rather than nephrogenic DI.

Patients with primary polydipsia frequently commence the test with low urine and plasma osmolalities due to water overload, and may fail to concentrate the urine adequately after the standard period of water deprivation since the plasma osmolality is frequently only increased into the low-normal range during this time. The renal response to desmopressin may also be impaired under these circumstances due to ‘washout’ of the renal concentrating gradients.

Hypertonic saline infusion

Severe DI can be diagnosed easily by a water deprivation test, but mild or borderline DI, and other more subtle defects of osmoregulation and thirst, may lead to equivocal findings despite protracted water deprivation. Under such circumstances, a hypertonic saline infusion (5% NaCl infused at 0.06 mL/kg per min for 120 min) may be used to delineate the pathology more accurately. Plasma osmolality and vasopressin are measured every 30 min but, as vasopressin measurements are not performed in most laboratories, this test is often only carried out in specialist centres. Discussion of the interpretation of this test is beyond the scope of this chapter (see Chapter 4 and Baylis, in Further reading, below).

Summary

Outline protocol for the investigation of a patient with pituitary disease

The multiple tests available and the controversies regarding their use and interpretation mean that the detail in which pituitary function is investigated remains a matter for clinical judgement. The authors propose the following protocol for the initial assessment of a patient with suspected pituitary disease.

1. Measure basal pituitary function on a single 09.00 h sample (see [Appendix 18.1, p. 371](#)).
2. If plasma cortisol is $<100\text{ nmol/L}$, then an IST is contraindicated and ACTH deficiency can be assumed (unless Addison disease is a possibility or the patient is receiving steroids).
3. Treat thyroid or ACTH deficiencies identified on basal concentrations (hypothyroidism reduces ACTH and GH responses to hypoglycaemia).
4. If plasma cortisol is $>100\text{ nmol/L}$ then:
 - a. if cortisol is $>400\text{ nmol/L}$, ACTH reserve can be assumed to be normal unless unexplained symptoms suggest otherwise
 - b. if cortisol is $<400\text{ nmol/L}$, perform a tetracosactide test – a response above 600 nmol/L confirms a normal axis (see above)
 - c. consider IST in the presence of equivocal responses to tetracosactide (e.g. peak 400–600 nmol/L), especially if there are unexplained symptoms, or if information on GH reserve is required.
5. Treat any ACTH deficiency identified (with hydrocortisone).
6. Consider whether investigation for GH deficiency is indicated:
 - a. if ACTH, thyroid and gonadal axes and osmolalities are normal, and the patient is asymptomatic with a normal AGHDA score, then assessment of GH reserve is probably of little *clinical* benefit (although it may be of *research* interest)
 - b. in the presence of other pituitary deficiencies, severe GH deficiency becomes increasingly likely with an increasing number of other deficiencies. If the AGHDA score suggests possibility of the clinical syndrome of adult GH deficiency, then perform a dynamic test of GH reserve, even if IGF-1 is normal

- c. at the initial assessment, or first assessment after pituitary surgery, many centres routinely perform an IST to formally document ACTH and GH reserve (except where contraindicated), even in the absence of symptoms.
7. Perform a water deprivation test if indicated by symptoms or by the basal osmolalities.

A clinical approach to assessment of the whole ACTH–adrenal axis

Assessment of the ACTH–adrenal axis is probably the most common assessment of complex endocrine function undertaken in clinical medicine, and remains the subject of ongoing controversy, variation in practice, differences in interpretation, artefact due to drugs and, frequently, the use of testing in inappropriate circumstances.

Assessment of adrenal reserve is broadly undertaken to answer the following four clinical questions.

1. Does this patient have undiagnosed hypoadrenalism? This question is frequently asked in the clinical context of hyponatraemia, hypotension and/or hypoglycaemia. A basal 09.00 h cortisol <100 nmol/L strongly suggests hypoadrenalism, and a value of >400 nmol/L makes it very unlikely (unless the patient is seriously unwell, in which case higher concentrations might normally be expected and the patient could be relatively hypoadrenal with diminished adrenal reserve).

When basal cortisol is uninformative, a short tetracosactide test is usually performed. A normal response (defined by an appropriate normal range – see above) effectively excludes adrenal deficiency. A definitely subnormal response confirms hypoadrenalism (as long as it is certain that the patient is not receiving steroid therapy), but should not immediately attract a diagnosis of Addison disease, since the prevalence of hypopituitarism is higher than that of primary adrenal failure.

2. Is the proven hypoadrenalism primary or secondary? In a newly diagnosed patient with hypoadrenalism, the clinical picture may contribute significantly to this differential diagnosis. Associated features may point towards primary adrenal failure: typical ‘Addisonian’ electrolytes (hyponatraemia, hyperkalaemia and slightly raised urea), typical ACTH-dependent hyperpigmentation or a personal or family history of autoimmune endocrine disease. Conversely, if the patient has visual field defects, clinical signs of hypogonadism or a pituitary mass, then ACTH deficiency is much more likely. Many patients, however, do not have clear-cut clinical clues to the diagnosis and a biochemical diagnosis is essential before embarking on life-long treatment.

Pragmatically, assessment of basal pituitary function is often the simplest and quickest approach, since it will either give evidence of other pituitary dysfunction, or, if normal, suggest primary adrenal failure (or, less likely, isolated ACTH deficiency).

Measurement of 09.00 h basal cortisol, ACTH and recumbent renin is typically slower, but diagnostic, since ACTH and renin are both elevated in primary adrenal failure.

The depot tetracosactide test has been used traditionally in this context (tetracosactide depot 1 mg i.m. with cortisol measured at intervals over 24 h). In primary adrenal failure, there is an absent or flat cortisol response, whereas in ACTH deficiency or HPA suppression by steroids, the cortisol response slowly rises to reach a near-normal peak at 24 h.

If primary adrenal failure is confirmed, measurement of adrenal antibodies may confirm autoimmune Addison disease, but if this is negative, alternative causes of adrenal destruction (tuberculosis, metastases) need to be considered and excluded. If ACTH deficiency is present, then full assessment of pituitary function and pituitary magnetic resonance imaging (MRI) are indicated.

3. Has this patient with known pituitary disease developed ACTH deficiency? This assessment has already been outlined above.

4. Does this patient on pharmacological steroid treatment still have adrenal suppression or could they now stop treatment? Numerically, this is probably one of the most frequent situations in which adrenal function is tested (with testing frequently being performed in inappropriate circumstances or the results being incorrectly interpreted):

- there is no point in performing any test of adrenal function when patients are taking supraphysiological doses of oral steroids (more than 5–7.5 mg prednisolone or 0.5 mg dexamethasone): basal cortisol and tetracosactide responses will always be suppressed
- high doses of inhaled, intranasal, topical or injected (for musculoskeletal disorders) steroids will also frequently suppress cortisol responses (basal concentrations may be undetectable). This often causes difficulties in interpretation, since most synthetic steroids are not measured in cortisol assays and it is impossible to differentiate biochemically between appropriate suppression and untreated deficiency
- prednisolone cross-reacts in many cortisol assays – so measurement of cortisol after a dose of prednisolone may give artefactually higher values. As a result of this, tests are often advised ‘under dexamethasone cover’. In the authors’ view, this is both unnecessary and unhelpful since the basal cortisol concentration (which gives useful information about background spontaneous cortisol secretion) will be undetectable. It is easier simply to omit prednisolone after 18.00 h on the day before testing, perform the relevant tests at 09.00 h and administer prednisolone when the test has been completed.

The authors therefore advise the following regimen when assessing whether or not pharmacological steroids can be discontinued:

- in a patient who is otherwise well, slowly reduce steroid doses to the lower end of the physiological replacement range indicated above before performing *any* hormonal tests

- then check the basal 08.00–09.00h cortisol one morning 24h after the last dose of steroid and before the steroid dose is taken that morning (tablets can be taken as usual immediately after the blood test). Interpretation:
 - <100 nmol/L: the patient still has profound adrenal suppression – further dose reduction would be unwise. Steroid replacement in the lower physiological range should be continued (if clinically well) and the test repeated in 6–12 months
 - >400 nmol/L: the adrenal axis is probably completely normal; steroids could be stopped (or slowly reduced if the underlying disease process contraindicates complete cessation immediately); further tests of adrenal axis are unnecessary unless there are unexplained symptoms
 - 200–400 nmol/L: adrenal function is not completely suppressed (and might be normal or partially suppressed); steroid doses can be slowly and cautiously reduced and the tests repeated
 - 100–200 nmol/L: significant adrenal suppression is more likely but will only remit if the steroid dose is reduced. Reduce steroid dose more slowly and cautiously; repeat 09.00h cortisol 2–3 months after each dose reduction.
- if concentrations are >300 nmol/L, and if the patient is clinically well, then steroid replacement can safely be stopped completely for several weeks in order to carry out a dynamic test of the adrenal using tetracosactide – otherwise the test may be indicated when on half-physiological dose as reassurance (it is unlikely to be normal, but profound suppression, for example a peak cortisol <300 nmol/L, would militate against further dose reduction). It may also be worth changing from a synthetic steroid to hydrocortisone, since this is shorter acting and therefore potentially less suppressive.

Even if the results of investigation are normal, it is important to remember that patients in whom the adrenal axis has been suppressed by long-term supraphysiological steroid administration may still require replacement during severe intercurrent illness for at least a year after steroids are discontinued.

Monitoring of pituitary function in disease states

Reassessment after pituitary surgery

Pituitary surgery, usually by the trans-sphenoidal route, is the recommended therapy for many pituitary adenomas. Following surgery, residual pituitary function requires careful monitoring and reassessment.

Transient DI is very common in the immediate postoperative period. However, the presence of nasal packs (required by the surgery) frequently results in a dry mouth so that the patient may drink excessive fluid, thereby causing polyuria in the absence of DI. Patients cured of acromegaly may also show polyuria in the absence of DI owing to rapid resolution of soft tissue enlargement. It is thus essential that DI is confirmed biochemically before replacement therapy is commenced. Paired urine and

plasma osmolalities should be checked daily for the first few postoperative days (and measured urgently if symptoms develop) and interpreted as outlined above. In borderline cases with persisting polyuria or nocturia, a formal water deprivation test may be required following recovery.

Most surgeons and physicians give glucocorticoid replacement therapy during the immediate perioperative period. One week postoperatively, or earlier if considered clinically appropriate and the patient is well, the evening replacement dose should be omitted and 09.00h plasma cortisol measured the following morning prior to the next dose. If a same-day cortisol result is available, the patient can remain off replacement, which should otherwise be restarted pending the result. A 09.00h cortisol <100 nmol/L indicates the need for continued, possibly life-long, replacement therapy. Patients with a 09.00h cortisol >300 nmol/L can be safely discharged without regular replacement therapy pending further investigations, but those with cortisol <400 nmol/L require steroid cover for intercurrent illnesses until the results are known. Patients with cortisol in the 100–300 nmol/L range usually need steroid replacement until definitive assessment is possible, but this must be determined by individual clinical assessment. In a retrospective review of our own practice, in which patients were investigated one week after pituitary surgery, 13% had a cortisol <100 nmol/L and 57% >400 nmol/L – so that only a minority of patients have equivocal responses at this stage.

Basal thyroid and gonadal function should be checked one week postoperatively. In view of the long plasma half-life of T₄, thyroid function should be reassessed at one month. At one week, a frankly low plasma testosterone in the male is usually diagnostic of ongoing gonadotrophin deficiency requiring replacement therapy. Lack of elevation of gonadotrophins in the postmenopausal female is also diagnostic. In both cases, borderline results require a repeat at one month. In the premenopausal female, early results may be misleading and it is wise to wait 2–3 months to assess the return of menstrual cyclicity prior to commencing replacement therapy.

A tetracosactide test is invalid as a test for ACTH reserve until at least two weeks postoperatively: the authors' practice is to perform this test approximately one month postoperatively in patients with a non-diagnostic basal cortisol concentration. The authors only perform a dynamic test of GH reserve in patients with symptoms suggestive of GH deficiency in whom replacement is being considered, but other centres prefer an IST at this stage to document fully postoperative pituitary function.

Monitoring after pituitary irradiation

Pituitary irradiation may be employed in the treatment of pituitary tumours (although its use is declining), and the hypothalamus and pituitary are also within the field of whole brain irradiation used for other diseases. All affected patients are at risk of slowly progressive hypopituitarism over at least the next 20 years and require regular monitoring. The order of progression is typically GH deficiency occurring early, LH and FSH deficiency next, and TSH and ACTH deficiency at a later stage. The lesion appears to be primarily at hypothalamic level and PRL deficiency therefore occurs rarely, if ever.

The development of GH deficiency in children is best assessed by careful monitoring of growth – and in adults, annual screening with the AGHDA questionnaire may be appropriate to identify patients in whom a dynamic test of GH reserve is required. For all other hormones, basal 09.00h pituitary function (see [Appendix 18.1, p. 371](#)) should be checked annually, with a subsequent tetracosactide test when serum cortisol is <400 nmol/L. (A routine annual tetracosactide test may actually prove to be a more efficient use of both the patient's and the laboratory's time.)

Monitoring in other pituitary disease states

Not all pituitary diseases result in progressive pituitary failure, so the need to monitor pituitary function in such patients is a matter of clinical judgement.

Patients with large pituitary mass lesions, granulomatous disease of the pituitary or hypothalamus, or unexplained acquired partial hypopituitarism certainly require monitoring of remaining pituitary function, which should be performed annually as above, at least for the first few years of follow-up. Conversely, patients with microprolactinomas whose PRL is suppressed on therapy, or patients with clearly defined isolated pituitary deficiency syndromes such as Kallman syndrome, are at minimal risk of progressive hypopituitarism and do not require regular monitoring in the absence of symptoms. Isolated GH deficiency in childhood is sometimes followed by other pituitary deficiencies in adult life, but it is as yet unclear whether this occurs sufficiently frequently to justify regular monitoring and, conversely, some such childhood cases are shown to have normal GH reserve when retested as young adults.

OTHER DIAGNOSTIC TECHNIQUES IN PITUITARY DISEASE

Clinical assessment

Endocrinologists continue to rely on clinical assessment to determine the need for further biochemical confirmation of diseases of the pituitary and hypothalamus. Investigations for possible Cushing syndrome and acromegaly are thus rarely performed unless they are suspected clinically because of the relevant symptoms and signs. Similarly, gonadotrophin deficiency is usually suspected by the development of loss of libido in both sexes, amenorrhoea in females and impotence and loss of secondary sexual hair in males. Diabetes insipidus results in thirst and polyuria, usually manifest in the earliest stages by nocturia (since urine is normally maximally concentrated overnight). Indeed, a biochemical nihilist might reasonably argue that there is no point in assessing these axes unless these clinical symptoms or signs are present.

Thyroid stimulating hormone deficiency may lead to typical symptoms and signs of hypothyroidism (see [Chapter 19](#)), although these are usually less severe than in primary thyroid disease. Adrenocorticotrophic hormone deficiency may result in general malaise, postural hypotension, headaches and abdominal pains and weight loss (pituitary cachexia). These symptoms of deficiency

are clearly non-specific and biochemical assessment is essential to confirm or deny deficiency. Growth hormone deficiency in adults results in impairment of quality of life in some patients – the pattern of symptomatology is non-specific, but can be recognized clinically or through testing with quality of life questionnaires.

Masses enlarging within the pituitary or hypothalamus frequently exert pressure on the optic chiasm and thus cause visual field loss, typically a bitemporal hemianopia. Clinical assessment of visual fields by confrontation and by formal charting is thus an essential part of the assessment of patients with pituitary disease. Large lateral extensions of pituitary tumours may more rarely result in cranial nerve deficits, the cavernous sinus syndrome or even temporal lobe epilepsy, and inferior extensions very rarely cause spontaneous cerebrospinal fluid rhinorrhoea.

Finally, panhypopituitarism is an unusual cause of unexplained hyponatraemia or hypoglycaemia and should always be considered during the investigation of these disorders.

Pituitary imaging techniques

Magnetic resonance imaging (MRI) scanning of the pituitary, with and without gadolinium enhancement, is the investigation of choice and allows detailed anatomical assessment. Computed tomography (CT) scanning of the pituitary gives less detailed information, but is used when MRI is contraindicated or impossible due to the presence of metallic implants or claustrophobia.

PITUITARY HYPERSECRETION STATES

Pituitary adenomas

Pituitary hypersecretion states are most usually associated with benign pituitary adenomas synthesizing and secreting the relevant hormone. Tumours may be subdivided into microadenomas (<10 mm diameter) and macroadenomas (>10 mm). Pituitary tumours may be invasive but are only very rarely malignant. Hyperplasia of the various cell types has also been described but, except when secondary to ectopic secretion of hypothalamic releasing hormones, is poorly defined. Non-functioning pituitary adenomas are more common in clinical practice.

Prolactinoma

Differential diagnosis of hyperprolactinaemia

A prolactinoma, or prolactin-secreting pituitary adenoma, is the commonest functioning pituitary adenoma. However, hyperprolactinaemia has many causes other than a prolactinoma ([Fig. 18.5](#)).

Any pituitary tumour or mass that is large enough to impede pituitary portal blood flow can result in hyperprolactinaemia by preventing hypothalamic dopamine from reaching the pituitary lactotrophs. Adenomas causing acromegaly also occasionally secrete PRL directly, in addition to GH. The differentiation of a prolactinoma from a functionless pituitary tumour may therefore sometimes be difficult, although biochemical results give some assistance. A plasma PRL >5000 mU/L

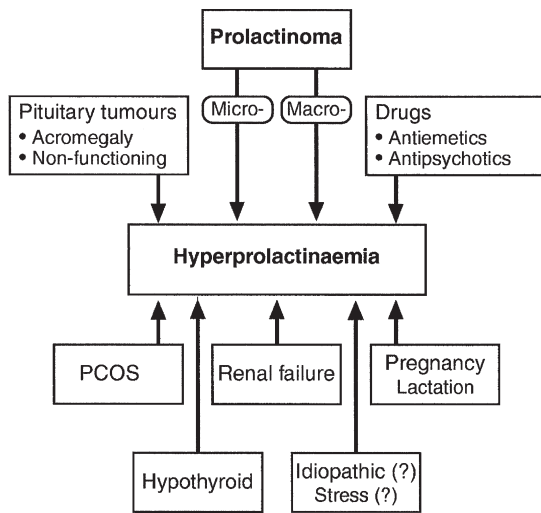


FIGURE 18.5 ■ Causes of hyperprolactinaemia. PCOS, polycystic ovary syndrome.

almost always indicates a prolactinoma, and concentrations may reach several hundred thousands. Plasma PRL <5000 mU/L may result from a prolactinoma, a functionless pituitary tumour (adenoma, craniopharyngioma etc.) or any of the other causes shown in [Figure 18.5](#). It is most unlikely for a large macroprolactinoma to be associated with PRL concentrations in this range. Conversely, while a small macroprolactinoma may be associated with an elevated PRL at any concentration, a small functionless microadenoma is unlikely to elevate plasma PRL into the thousands. Thus, when PRL is raised, but <5000 mU/L, the smaller the adenoma seen on CT or MRI scan, the more likely it is to be a prolactinoma and vice versa.

Any drug with dopamine antagonist effects may cause hyperprolactinaemia; this includes almost all antiemetics (e.g. metoclopramide, prochlorperazine, domperidone) and most major tranquilizers (chlorpromazine and other phenothiazines, haloperidol, risperidone etc.). A careful drug history is therefore essential in the assessment of hyperprolactinaemia. Antiemetics can often be discontinued but antipsychotic medication usually needs to be continued, although change to a newer antipsychotic with less effect on PRL may be possible (e.g. quetiapine, olanzapine and other ‘atypical antipsychotics’).

Primary hypothyroidism may be associated with hyperprolactinaemia in the low thousands (mU/L) (and with galactorrhoea and oligo- or amenorrhoea) that resolves with thyroid replacement therapy. Chronic kidney disease causes hyperprolactinaemia of a similar degree. Measurement of thyroid and renal function is thus essential in any patient with hyperprolactinaemia.

Mild and often intermittent hyperprolactinaemia (usually <1000 mU/L) is often seen in polycystic ovary syndrome (PCOS). In the appropriate clinical context (hirsutism, acne, obesity), assessment of plasma androgen concentrations and sex hormone-binding globulin (SHBG) may therefore be indicated (see [Chapter 22](#)). Major ‘stress’, for example hypoglycaemia, also elevates plasma PRL but it is far from clear whether psychological stress or fear of venepuncture is able to do so.

Finally, it should be remembered that by far the commonest causes of a state of hyperprolactinaemia and amenorrhoea are pregnancy and lactation: a pregnancy test may be indicated.

Role of dynamic tests of PRL secretion

In order to assist the differential diagnosis of hyperprolactinaemia, a number of dynamic tests of PRL secretion have been advocated and used, particularly in continental Europe. Such tests include the TRH test, metoclopramide test and domperidone test. None of these tests gives clear differentiation between prolactinomas and other causes of hyperprolactinaemia and they have no role in routine biochemical assessment.

Assessment of remaining pituitary function

Few patients with small macroprolactinomas have deficiencies of other pituitary hormones, but, conversely, a functionless macroadenoma can cause severe hypopituitarism associated with mild hyperprolactinaemia. Clearly not every patient with a high PRL requires full pituitary function testing. The authors’ own (arbitrary) practice is to perform basal anterior pituitary function tests and a pituitary MRI whenever the plasma PRL is consistently >1000 mU/L, but not otherwise unless there are relevant symptoms or signs.

Outline of presentation and management of prolactinoma

In females, hyperprolactinaemia causes oligo- or amenorrhoea, galactorrhoea and breast tenderness, poor libido and infertility. Patients may present with any combination of these symptoms, with or without other signs of pituitary enlargement such as visual field defects. In males, poor libido and impotence are the only early symptoms and patients frequently present at a later stage with visual field loss, with or without headaches. Development of a prolactinoma in the peripubertal period can also result in delayed or arrested puberty in both sexes, and a serum PRL measurement is thus an essential part of the assessment of such patients.

The management of prolactinomas is primarily medical, using dopamine agonist therapy, which can suppress plasma PRL to normal and cause clinical remission of all symptoms. The long-acting dopamine agonist cabergoline is now most commonly used – bromocriptine and quinagolide are alternatives. Bromocriptine may be preferred when pregnancy is planned, in view of its longer safety record, although there is no evidence of fetal harm with the newer drugs. In addition to lowering PRL and relieving symptoms, dopamine agonist therapy also causes dramatic shrinkage of macroprolactinomas in a majority of cases, with resolution of pressure effects such as visual field loss. Such tumour shrinkage is the ultimate non-invasive method to distinguish a prolactinoma from other pituitary lesions causing hyperprolactinaemia. In a majority of prolactinomas the hyperprolactinaemia will recur if dopamine agonist therapy is stopped, so that treatment is usually long term, although a significant minority may be able to discontinue therapy.

Alternative therapies for a prolactinoma include trans-sphenoidal surgery and, rarely, pituitary irradiation.

Monitoring the response to dopamine agonist therapy

The aim of dopamine agonist therapy is to suppress plasma PRL concentrations into the normal range. The dose of dopamine agonist is usually slowly increased with sequential PRL estimations until this is achieved. Thereafter, plasma PRL would normally be measured at each clinic visit, although these need only be annual in patients on long-term therapy. Fibrotic reactions have been described as rare side-effects of long-term dopamine agonist therapy (pulmonary, retroperitoneal and pericardial); some form of monitoring is advised and the authors' practice is to monitor for symptoms and measure urea and electrolytes, full blood count and C-reactive protein (CRP) (or plasma viscosity or erythrocyte sedimentation rate (ESR)) annually.

Macroprolactinaemia

In some patients, PRL circulates bound to immunoglobulin forming an immunoreactive PRL of high molecular weight (macroprolactin). This is biologically inactive but in molar terms may considerably exceed the concentration of free PRL. Commonly used immunoassays for PRL cross-react with macroprolactin to a variable extent, and with some assays macroprolactin is a common cause of hyperprolactinaemia. This can usually be detected by polyethylene glycol (PEG) precipitation, although gel filtration chromatography is more reliable. Laboratories should know the susceptibility of their own assay to this artefact, and always arrange to screen for macroprolactinaemia if persistent hyperprolactinaemia does not correlate with clinical symptoms or expected responses to treatment.

Hook effect

Extremely high PRL concentrations (say >100 000 mU/L) in patients with large pituitary macroprolactinomas may, rarely, exceed the usual linear range of the assay calibration curve when homogeneous assays are used, and give apparently only modest elevations not suggestive of a large prolactinoma. This can be detected by assays of doubling dilutions of serum (when measured PRL concentration fails to fall as expected and may begin to rise).

Acromegaly

Diagnosis of acromegaly

The diagnosis of acromegaly is usually suspected clinically before any biochemical tests are performed. However, by the time acromegaly is clinically obvious, the patient may have already suffered considerable morbidity from soft tissue changes, arthritis, hypertension, diabetes and consequent cardiovascular disease. Because of this, the aim should be to diagnose acromegaly at an early stage, when clinical signs are less clear cut and differentiation from normality can only be confirmed by biochemical investigations.

Basal GH and IGF-1 estimation. Random plasma GH measurements are never diagnostic of acromegaly, since high concentrations may occur with normal pulses of secretion and with stress. Nevertheless, a low plasma GH (<0.4 µg/L), together with a normal IGF-1 essentially excludes the diagnosis, whereas a high IGF-1 with consistent GH concentration makes the diagnosis very likely. It is therefore the authors' practice to measure a random GH and IGF-1 concentration as a simple first test when acromegaly is suspected. An age-related normal range for IGF-1 must be used for interpretation since concentrations vary considerably throughout life.

Glucose tolerance test. The 75 g glucose tolerance test (GTT) remains the gold standard test for the diagnosis of acromegaly. After glucose, plasma GH in normal individuals is suppressed to <0.4 µg/L. Patients with acromegaly fail to show complete suppression and may show a paradoxical rise of plasma GH. Failure of GH to fall to <0.4 µg/L is highly suggestive of acromegaly, even if substantial suppression of GH has occurred, and an elevated IGF-1 will normally confirm the diagnosis.

Failure of normal glucose suppression of GH may occur in the absence of acromegaly in liver disease, poorly controlled diabetes, chronic kidney disease, malnutrition (including anorexia nervosa) and primary growth hormone insensitivity (Laron syndrome). Diabetes can be a particular diagnostic challenge, since a GTT is inappropriate in established disease. Insulin-like growth factor 1 concentrations, however, are typically normal in uncomplicated diabetes, and, if elevated, the authors usually recommend a four-point GH day curve to confirm circulating GH concentrations consistently above the 'safe' range.

Other diagnostic tests. Patients with acromegaly may show a paradoxical rise in plasma GH during standard TRH or LHRH tests. When acromegaly has been clearly diagnosed by other means, these tests give no additional clinically useful information, but they may help confirm mild acromegaly in patients when the results of other tests are equivocal.

In normal individuals, plasma GH shows a transient rise during a dopamine infusion (4 µg/kg per min for 3 h), whereas patients with acromegaly characteristically show a fall in plasma GH. This test can also be used to confirm an equivocal diagnosis but is not used routinely.

Ectopic secretion of GHRH, for example from a pancreatic endocrine tumour, is a rare cause of clinical acromegaly. The diagnosis is not usually suspected until somatotroph hyperplasia is found at trans-sphenoidal surgery rather than the expected adenoma. In these circumstances, peripheral plasma GHRH concentration (available only in a few specialist centres) is elevated and diagnostic.

Outline of presentation and management of acromegaly

Acromegaly is usually diagnosed clinically when the patient presents with an incidental illness to a doctor who recognizes the typical physical appearance, which is frequently longstanding and only slowly progressive.

Growth hormone-secreting tumours that develop in childhood result in gigantism, but after puberty, fusion of the major epiphyses means that growth in height is no longer possible. Patients with acromegaly do, however, develop excessive growth of soft tissues and small bones of the skull, face, hands and feet, which results in the characteristic clinical features. They have large, 'spade-like' hands and large, broad feet, usually with a history of increasing ring size and shoe size over a number of years. Facial appearance is characteristic with prognathism, prominent supraorbital ridges, large nose and tongue and generalized coarsening of the features. Examination of old photographs will frequently demonstrate that these changes have been progressive over 5–10 years or more. Headaches, often severe, and troublesome, inappropriate sweating are common symptoms; other findings frequently include a multinodular goitre, kyphosis, hypertension and diabetes mellitus, the last two presumably contributing to the observed increase in cardiovascular morbidity and mortality in this condition.

Patients with acromegaly often develop deficiencies of other pituitary hormones, and full investigation of pituitary function is mandatory once the diagnosis has been made.

Although acromegaly frequently goes undiagnosed for many years, treatment is almost always indicated since, if left untreated, the condition results in disfigurement, morbidity (primarily due to kyphosis and osteoarthritis) and mortality (primarily cardiovascular, but possibly also an increase in malignancies).

First-line therapy in almost all patients is transphenoidal surgery, which is able to cure at least 75% of microadenomas but a much smaller proportion of large macroadenomas, some of which are locally invasive. A number of studies have shown substantial differences in success rates between surgeons and centres, and an experienced pituitary surgeon is essential. Unsuccessful pituitary surgery is traditionally followed by pituitary irradiation, which steadily lowers GH secretion, but may require 10–20 years to normalize GH concentrations in patients where the initial GH is very high (over 40 µg/L). Medical therapy with dopamine agonists such as high-dose cabergoline (e.g. 0.5 mg daily) is useful in a proportion of patients. The potent, long-lasting somatostatin analogues octreotide and lanreotide are able to normalize GH concentrations in the majority of patients (but not all), albeit at the expense of monthly injections and a high economic cost. Both classes of drugs may allow significant shrinkage of a somatotroph adenoma preoperatively. More recently, the GH antagonist pegvisomant has been shown to be capable of normalizing IGF-1 in almost every patient, but treatment requires daily injections and is extremely expensive. The success of these drugs has led some authorities to recommend long-term medical therapy instead of radiotherapy (or even surgery), but this is not currently routine practice in most centres.

Monitoring the response to therapy

There has been considerable controversy over the years regarding the appropriate biochemical criteria to diagnose remission or cure of acromegaly after therapy, but results of large retrospective studies and consensus

meetings have achieved much greater uniformity. The current consensus criteria for disease control in acromegaly define active and controlled disease. The criteria for active disease are a random GH >1 µg/L *and* nadir GH after OGTT ≥0.4 µg/L with an elevated IGF-1. Patients with active disease usually require closer monitoring to ensure the clinical and biochemical picture does not deteriorate, and would be considered for additional treatment if the clinical picture were to deteriorate or the GH increases to >2 µg/L. The criteria for controlled disease are a random GH <1 µg/L *or* nadir GH after OGTT <0.4 µg/L and an age-sex normalized IGF-1 for age and sex. Occasionally after treatment for acromegaly there is a discrepancy between GH and IGF-1 concentrations, most commonly involving a normal GH and elevated IGF-1 concentration. Although the aim of management should be to normalize both GH and IGF-1, it is currently felt that a normal GH level concentration is probably the more important determinant of cardiovascular risk reduction. The current literature suggests that a GH <2 µg/L is probably within the safe range for cardiovascular disease, although there remains debate about whether adjuvant treatment is indicated for patients who have discordantly elevated IGF-1 despite acceptable growth hormone concentrations.

Diagnosis and differential diagnosis of Cushing syndrome

Endogenous Cushing syndrome is rare, with an incidence of approximately 5 per million population per year. However, many of the clinical features of the syndrome are very common (weight gain, hypertension, hirsutism, diabetes) so that biochemical tests are frequently requested to exclude the diagnosis and are essential to confirm its presence. Clinical features certainly increase the suspicion of Cushing syndrome (easy bruising, thin skin, red/purple abdominal striae, proximal myopathy, specific distribution of fat, e.g. supraclavicular and dorsal fat pads in addition to central obesity). However, many patients with biochemically and surgically proven Cushing disease are not as grossly 'cushingoid' as seen in a patient on high-dose pharmacological steroids or as illustrated in older medical student textbooks; they are often overweight rather than obese but with a history of unexpected recent weight gain and change towards central distribution of fat.

Measurement of urinary free cortisol (UFC) is frequently used as a first-line screening test for Cushing syndrome. It has the advantage of simplicity, but the disadvantage that the diagnosis may be missed if urine collection is incomplete, and of false positive elevations of urinary free cortisol in obesity and stressful illness.

Many clinicians will therefore use some form of low-dose dexamethasone suppression test as their preferred first-line screening test.

Low-dose dexamethasone suppression test. Dexamethasone 0.5 mg is administered 6-hourly for 48 h commencing at 09.00 h, with plasma cortisol measured at 09.00 h at the end of the test. Normal response is a plasma cortisol <50 nmol/L (and undetectable in most normal individuals). Many regard this test as too

complex – but it still potentially involves only one blood sample and one prescription, and patients can easily follow the administration protocol as an outpatient if given a simple information sheet. It is sensitive and much more specific than the overnight test and is the authors' preferred screening test.

Overnight dexamethasone suppression test.

Dexamethasone 1 mg is administered on going to bed at 23.00 h and plasma cortisol is measured at 09.00 h. Suppression of cortisol to <50 nmol/L is unequivocally normal and because it is a simple and sensitive investigation, it is preferred by many clinicians, although it has a higher false positive rate than a formal 48 h test.

Midnight salivary cortisol. This test is gaining increasing popularity for diagnosis of Cushing syndrome, particularly in the USA. In normal individuals, plasma cortisol concentration is low at midnight, when resting or asleep, whereas in Cushing syndrome, the circadian rhythm is lost or reduced. A midnight sleeping plasma cortisol <100 nmol/L has traditionally been used to exclude Cushing syndrome, but the practicalities of inpatient bed and junior staff availability mean that this test is now rarely practical in the UK. More recent data suggest that an awake value >207 nmol/L may have a higher specificity and is easier than sampling whilst the patient is asleep. A saliva sample can be collected by the patient at home and midnight salivary cortisol has been shown to be a useful screening test, although the precise cut-off value depends upon the local reference range and the assay is not yet widely available in clinical practice.

Other tests. All tests for diagnosis and differential diagnosis are subject to false positive and false negative results. Where the result of first-line screening tests are in conflict with the clinical findings, or where doubt exists, then other tests are sometimes performed: repeated UFC measurements to establish the presence or absence of 'cyclical Cushing syndrome'; a formal 'circadian rhythm' assessment of plasma cortisol as an inpatient; an insulin stress test (IST) to demonstrate suppression of the cortisol response to stress, which is typically seen in Cushing syndrome. The desmopressin stimulation test, where patients with pituitary Cushing disease typically show a response, and normal individuals do not, is no longer felt to be sufficiently useful for routine clinical practice.

At this stage, advice from a specialist tertiary centre is usually required. Scanning of the adrenals or pituitary is not advised until the diagnosis of Cushing syndrome is clearly established, since 'incidentalomas' of both the pituitary and adrenals are common and may lead to an inappropriate diagnosis.

Once Cushing syndrome has been diagnosed, measurement of plasma ACTH can readily distinguish primary adrenal causes from ACTH-dependent Cushing syndrome. The former can be readily confirmed on CT or MRI scan of the adrenals, but the differential diagnosis of the latter remains a challenging clinical problem (see Fig. 18.6).

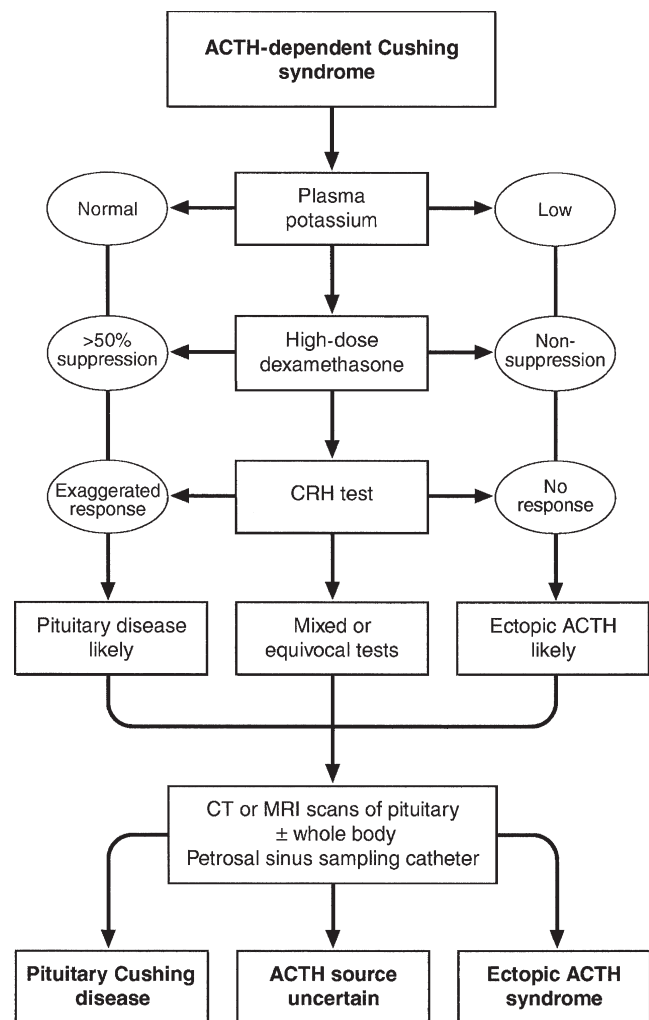


FIGURE 18.6 ■ Strategy for the differential diagnosis of ACTH-dependent Cushing syndrome.

Clinical context of ACTH-dependent Cushing syndrome

Adrenocorticotrophic hormone-dependent Cushing syndrome may be due either to pituitary Cushing disease, usually associated with a corticotroph microadenoma, or to ectopic ACTH secretion (and very rarely ectopic CRH secretion) by a variety of tumours elsewhere.

Numerically, the commonest source of ectopic ACTH secretion is small cell carcinoma of the bronchus. In this case, the clinical presentation is usually rapid and is often with symptoms directly related to the primary tumour. Although plasma ACTH and cortisol concentrations can be extremely high, such patients are rarely clinically cushingoid but may have pigmentation (due to the ACTH), proximal myopathy and hypokalaemic alkalosis. Such patients rarely present diagnostic difficulties.

In contrast, other, often relatively benign neuroendocrine tumours, can also secrete ACTH, particularly carcinoid tumours of the lungs (the commonest), mediastinum and elsewhere; pancreatic endocrine tumours; medullary thyroid carcinomas and, rarely,

phaeochromocytomas. Such tumours are small, slow growing and clinically occult and these patients may present with typical Cushing syndrome indistinguishable from that seen in pituitary-dependent disease. Such cases represent 10–15% of ACTH-dependent Cushing syndrome in most large series. Since these tumours may be only a few millimetres in diameter and thus hard to localize clinically or by conventional imaging techniques, biochemical tests are of paramount importance in their differentiation from pituitary Cushing disease.

Plasma cortisol and ACTH concentrations

In pituitary disease, plasma ACTH is often <100 ng/L, but higher values (up to at least 400 ng/L) are not infrequently seen. While, on average, plasma ACTH (usually >100 ng/L), and thus plasma cortisol, are higher in occult ectopic ACTH secretion than in pituitary disease, there is overlap and values usually lie within the range observed in Cushing disease.

Chromatography of plasma ACTH has been employed in the past to identify large molecular weight ACTH precursors, which are frequently secreted ectopically by tumours, but also occasionally by pituitary tumours. The complexity of these techniques means that they are no longer generally available, although immunometric assays specific for ACTH precursors may sometimes provide equivalent information.

High-dose dexamethasone suppression test

The high-dose dexamethasone suppression test was initially described in 1960 for the differentiation of adrenal from pituitary causes of Cushing syndrome. While largely superseded in this role, it remains useful in the differentiation of occult ectopic ACTH secretion.

A number of dosage regimens have been described, but the authors prefer the original regimen, with which there is greatest experience. Dexamethasone is administered orally, 2 mg strictly six-hourly for 48 h, commencing at 09.00 h. Plasma cortisol is measured at 09.00 h (basal value) and after 24 h and 48 h of treatment. 'Significant suppression' is arbitrarily defined as a fall in plasma cortisol to 50% or less of the basal value.

Using such criteria, about 80% of patients with Cushing disease, but only 10% of patients with occult ectopic ACTH secretion, show significant suppression of plasma cortisol. Thus differentiation is far from absolute, but since pituitary disease is far more common this means that significant suppression during the test statistically favours pituitary disease by approximately 70:1, whereas patients showing no suppression at all are equally likely to have a pituitary or ectopic source.

Corticotrophin releasing hormone test

Following the isolation and synthesis of the 41-amino acid CRH, the CRH test (see [Appendix 18.1, p. 372](#)) has gained an important role in the differential diagnosis of ACTH-dependent Cushing syndrome.

Following administration of CRH (in the UK, usually human CRF-41, but in the USA, typically ovine CRF-41, in both cases 100 µg i.v.), most patients with pituitary Cushing disease show a rise in plasma ACTH and therefore cortisol (defined as an increase of 25% above basal). In about 80% of patients with pituitary disease, the rise in plasma cortisol is 'exaggerated', that is, above the peak value observed in normal individuals. It is of interest and of practical importance that patients who fail to suppress with high-dose dexamethasone usually do show a cortisol response to CRH. In contrast, a significant cortisol response to CRH in ectopic ACTH secretion is extremely rare and limited to one or two isolated case reports in the world literature.

Other tests

An unprovoked hypokalaemic alkalosis remains an important clue to the presence of ectopic ACTH secretion, since over 90% of such patients present with hypokalaemia, compared with only 10% of patients with pituitary Cushing disease.

In the context of ectopic ACTH, tumours frequently also secrete other peptide hormones so that one or more tumour markers may be positive in the blood of such patients. It is therefore the authors' practice in patients with equivocal tests to screen plasma for tumour markers including calcitonin, human chorionic gonadotrophin (hCG), α -fetoprotein, carcinoembryonic antigen (CEA) and 'gut hormones', any of which may occasionally prove to be positive.

Petrosal sinus sampling for ACTH

It can be seen from the discussion above that no test is completely reliable in the differential diagnosis of ACTH-dependent Cushing syndrome. Because of this, techniques have been developed to ascertain directly the source of ACTH secretion via bilateral, simultaneous, inferior petrosal sinus sampling (IPSS) for ACTH after CRH stimulation.

Venous sampling catheters are passed from both femoral veins to lie in the respective inferior petrosal sinuses, which drain blood from the cavernous sinus and pituitary gland. The patient is partially heparinized prior to the final placement of the catheters to avoid the risk of venous thrombosis affecting the brain, which has been reported rarely. Blood for ACTH estimation is obtained simultaneously from both sides and from a peripheral vein, at timed intervals before and after an intravenous bolus of CRH. It is the authors' practice to administer TRH (200 µg) at the same time to measure central secretion of TSH and PRL to confirm the equivalence of catheter placement since, in the authors' experience, unequal sampling is frequently seen despite good radiological appearances.

After a technically successful procedure, the vast majority (all in most series) of patients with pituitary Cushing disease show a central:peripheral gradient of ACTH concentration of >2:1 basally or >3:1 after CRH, which is diagnostic of pituitary disease. Significant lateralization of ACTH secretion may also be observed and may guide

the surgeon and improve the surgical cure rate. It is the authors' practice to perform IPSS in almost every patient with ACTH-dependent Cushing disease without an obvious ectopic tumour or pituitary macroadenoma, as part of the routine differential diagnostic work-up; others reserve the test for situations where other tests are contradictory. The relative complexity of the technique means, however, that it should be restricted to those centres with experience in its use.

Venous sampling for ACTH from a wide variety of sites throughout the body, using a catheter in the central veins, occasionally may also give useful information, especially when directed towards a suspicious lesion detected on body CT, MRI or positron emission tomography (PET) scanning.

Imaging

Although imaging plays an important role in the differential diagnosis of Cushing syndrome, it is important to interpret the biochemical findings as the imaging may be misleading. An MRI scan of the pituitary is frequently normal in pituitary Cushing disease, and may be abnormal in 10% of normal individuals (and therefore in ectopic ACTH patients). In some patients, where the suspicion of ectopic ACTH secretion is high on biochemical grounds, detailed CT scanning of the entire thorax and abdomen may be necessary. An appropriate diagnostic strategy is summarized in [Figure 18.6](#).

Outline of management

The first-line treatment of Cushing disease is now selective adenectomy at trans-sphenoidal surgery, which results in cure in about 75% of cases, varying from series to series and with the biochemical criteria used to diagnose 'cure' (see below). The earlier alternatives of bilateral adrenalectomy or pituitary irradiation are now usually reserved for patients who are not cured by pituitary surgery.

Medical therapy is also possible for Cushing disease of all types, but is usually reserved for preoperative preparation or in patients where surgery has failed or is impossible. Commonly used drugs include metyrapone and ketoconazole. The aim of therapy is to maintain mean circulating cortisol concentrations between 150 and 300 nmol/L, but care must be taken to ensure that the cortisol assay used does not cross-react with precursors above the drug-induced block in cortisol synthesis (11-deoxycortisol in the case of metyrapone).

Reassessment after pituitary surgery

As with acromegaly, definition of 'cure' after trans-sphenoidal surgery for Cushing disease remains a controversial area. Many of the larger series reporting success of this type of surgery have failed to define their criteria for 'cure', and some appear to be assessed only by clinical remission or return of UFC to within the reference range.

After successful pituitary surgery, most patients have adrenal insufficiency, since the normal

hypothalamo-pituitary axis is fully suppressed by the longstanding hypercortisolaemia. To be confident of 'cure', a 09.00 h cortisol, off replacement therapy, at the early preoperative reassessment should be subnormal, that is, <100 nmol/L (some authors insist on undetectable concentrations). Other authors suggest that a concentration <140 nmol/L in the context of clinical improvement may suggest remission. Occasionally, delayed remission may occur, although a basal cortisol >140 nmol/L after six weeks post-pituitary surgery is an indication that further treatment is needed. In cases of doubt, a formal 48 h low-dose dexamethasone suppression test should be performed (0.5 mg, orally, strictly six-hourly for 48 h, commencing at 09.00 h); normal individuals suppress plasma 09.00 h cortisol to undetectable values (<25 nmol/L in the authors' laboratory) at the end of this test, and failure to show such complete suppression indicates residual disease, albeit mild in many cases.

Patients in remission from Cushing disease require glucocorticoid replacement therapy for several months or even years. To assess the possibility of steroid withdrawal, a 09.00 h cortisol should be measured periodically, having omitted the early morning replacement dose and the evening dose the day before. When cortisol rises to >100 nmol/L, the replacement dose can be slowly reduced, with continued monitoring of cortisol concentrations and a tetracosactide test or insulin stress test (IST) performed once replacement is completely withdrawn. Patients successfully withdrawn from replacement therapy require periodic monitoring, with UFC and/or low-dose dexamethasone suppression testing to detect early evidence of relapse.

Thyroid stimulating hormone-secreting adenomas

Thyroid stimulating hormone-secreting pituitary adenomas (TSHomas) are rare, but occasional cases will be encountered in most centres. Patients typically present with mild clinical and biochemical thyrotoxicosis with an inappropriately 'normal' or slightly elevated plasma TSH.

The differential diagnosis of a patient with a raised free T4 and normal or elevated plasma TSH should include thyroid hormone resistance (THR), and assay interference. Although only patients with TSHomas should be clinically hyperthyroid, it is often difficult to distinguish TSHomas from THR or assay interference on clinical grounds alone, and other tests are often needed. Patients with THR typically show a rise in TSH after injection of TRH, whereas a flat response is usually seen with a TSHoma, aiding the differential diagnosis. The α -subunit (of TSH) can also be measured by some centres and is typically elevated with a TSHoma but not with THR. Analysis of the thyroid hormone receptor gene (or more simply measuring thyroid function in family members) may clinch the diagnosis of THR. Measurement of sex hormone binding globulin (SHBG) is helpful in the differential diagnosis because SHBG tends to be normal in the thyroid hormone resistance and elevated in the

TSHoma due to the increase in SHBG produced by the liver in response to thyrotoxicosis.

Treatment of TSHoma is by somatostatin analogue therapy and/or pituitary surgery, with or without pituitary irradiation and carbimazole therapy.

Gonadotrophin-secreting adenomas

Clinically functioning gonadotrophin-secreting adenomas of the pituitary are extremely rare. In the male, they present with elevated plasma testosterone, usually with enlarged testes, in the presence of 'normal' or elevated LH and FSH. In the female, the condition may mimic premature ovarian failure, with amenorrhoea and elevated gonadotrophins. Of theoretical interest, however, is the experimental observation that a majority of so-called functionless pituitary adenomas are able to synthesize and secrete LH, FSH or their common α -subunit *in vitro*, the significance of which is unclear.

HYPOTHALAMIC AND PITUITARY DEFICIENCY STATES

Diseases that may lead to generalized hypopituitarism

Non-functioning pituitary adenomas

Pituitary adenomas that do not result in the hypersecretion syndromes described above are described as non-functioning or functionless pituitary adenomas. As mentioned above, these can frequently be shown experimentally to be capable of synthesis of LH and FSH and occasionally of any other pituitary hormone, in subclinical amounts. Such patients usually present either with symptoms and signs of hypopituitarism or with other signs of pituitary tumour enlargement, primarily visual field loss. Treatment is by pituitary surgery, followed by irradiation where appropriate. Although such adenomas do not show dramatic shrinkage in size with dopamine agonist therapy, recent evidence suggests that these drugs may reduce the risk of recurrence after successful surgery (or during conservative observation with smaller tumours).

Other pituitary and parasellar tumours

Craniopharyngiomas, derived from remnants of Rathke's pouch, are the only other common type of primary pituitary tumour. These usually present in childhood with poor growth, hypopituitarism and signs of pituitary enlargement, but may also present for the first time in adult life. Malignant tumours from elsewhere in the body occasionally metastasize to the pituitary; in addition to the usual signs of pituitary tumour, severe headache is a frequent symptom.

A wide variety of other tumours may also present with symptoms and signs of a pituitary or hypothalamic mass including hypothalamic germinomas, meningiomas, optic nerve and other gliomas, chordomas, hamartomas and pineal tumours.

Inflammatory diseases and disorders of unknown aetiology

Inflammatory conditions involving the hypothalamo-pituitary axis are rare and typically present with an unusual pattern of hypopituitarism in conjunction with diabetes insipidus. Sarcoidosis may cause granulomatous masses in the pituitary and hypothalamus, with or without evidence of disease elsewhere; measurement of plasma angiotensin-converting-enzyme concentration may, therefore, occasionally be helpful in the diagnosis of unexplained pituitary masses. Other granulomatous diseases that may present in the same way include tuberculosis and histiocytosis X.

Lymphocytic hypophysitis is an inflammatory disorder of unknown but possibly autoimmune aetiology; it occurs most frequently during pregnancy or the puerperium, but may occur at other times and less frequently in the male. Granulomatous hypophysitis is probably part of the same spectrum of diseases. Idiopathic pituitary fibrosis is also rarely seen and may be associated with other midline fibrosis syndromes. All these conditions may present with hypopituitarism or with signs of a pituitary mass.

Other conditions

Sheehan syndrome is hypopituitarism due to pituitary infarction secondary to massive postpartum haemorrhage and usually presents with failure of lactation and amenorrhoea in the puerperium. New cases should be rare with modern obstetric management, but some patients continue to present in later life, usually with a history of amenorrhoea since the birth of the last child, sometimes 30–40 years previously; diagnosis of such patients is usually precipitated by intercurrent illness.

Pituitary apoplexy is caused by pituitary haemorrhage or infarction, often associated with a previously undiagnosed pituitary adenoma. Patients may present with headache, visual loss, symptoms of hypopituitarism and, sometimes, meningism due to the presence of blood in the subarachnoid space; an initial working diagnosis is often meningitis or sub-arachnoid haemorrhage. Surgical decompression is frequently required. Head injury, especially when associated with skull fractures through the pituitary fossa, may also be associated with hypopituitarism, including diabetes insipidus. Congenital developmental disorders, such as septo-optic dysplasia, may also have hypopituitarism as part of a larger syndrome.

Finally, a substantial number of patients with isolated and generalized hypopituitarism have 'idiopathic' disease, with no lesion demonstrable on CT scanning. These presumably represent selective failure of the hypothalamic neurons that secrete the relevant releasing hormone. At least some of these cases may be autoimmune in origin, but more research is required to define their precise aetiology.

Growth hormone deficiency

Growth hormone deficiency most frequently arises in the context of other pituitary diseases discussed above. Isolated, idiopathic GH deficiency is, however, also

relatively common and usually presents with short stature and poor growth in childhood. The most important feature of the management in childhood is accurate measurement of growth and growth velocity, usually followed by assessment of GH reserve if growth is falling away from the third centile on standard charts. Treatment of established GH deficiency is with biosynthetic human GH (somatotrophin).

There are, however, many other causes of short stature in the absence of GH deficiency. These include hypothyroidism, chronic kidney disease, malabsorption, Crohn disease and other inflammatory disorders (which may be otherwise asymptomatic), asthma, congenital heart disease, Turner syndrome and other congenital disorders, and familial short stature. Laboratory investigations clearly have a major role in excluding a number of these possibilities.

Primary growth hormone insensitivity (Laron syndrome) is a rare congenital cause of severe short stature caused by absence or mutation of the GH receptor. Circulating GH concentrations are high and nonsuppressible, in the presence of low or absent circulating IGF-1 concentrations. Therapy with recombinant IGF-1 is possible.

Growth hormone deficiency in adults was previously considered to be of no significance. However, many studies have now shown that treatment of such adults with somatotrophin may improve quality of life, decrease body fat and increase bone density and muscle mass and strength. In the UK, the National Institute for Health and Care Excellence (NICE) have recommended replacement therapy under strictly defined circumstances.

Gonadotrophin deficiency

Gonadotrophin deficiency may present at birth with an undervirilized male, with failure of pubertal development or with loss of previously normal gonadal function. Isolated gonadotrophin deficiency may be idiopathic, but is sometimes familial and associated with anosmia (Kallman syndrome) and sometimes cleft lip or palate.

Treatment of gonadotrophin deficiency is with the relevant gonadal steroids, testosterone or oestrogen, given by mouth, injection, implant or transdermally. If fertility is desired, complex regimens involving gonadotrophin therapy or pulsatile subcutaneous administration of GnRH are required.

Gonadotrophin deficiency is diagnosed by the combination of low plasma gonadal steroid concentrations with low or inappropriately 'normal' concentrations of LH and FSH. However, a large number of other conditions may also produce this biochemical pattern in the absence of organic gonadotrophin deficiency, particularly in the female (see Fig. 18.7).

Interpretation of borderline testosterone concentrations

A common clinical and biochemical dilemma is the interpretation of an isolated testosterone that is near or below the lower end of the reference range (say 7.5–12 nmol/L), with normal gonadotrophins, in the presence of commonly occurring symptoms that might indicate hypogonadism (e.g. low libido or erectile dysfunction). The

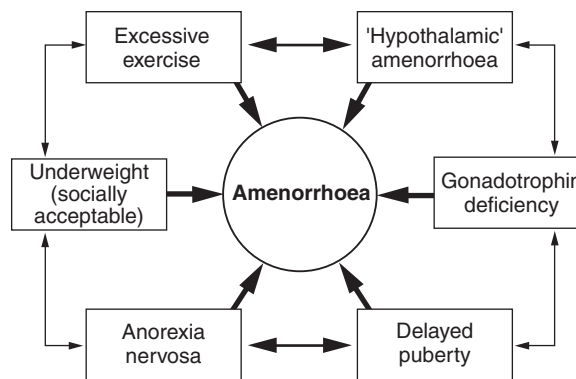


FIGURE 18.7 ■ Inter-related conditions causing amenorrhoea and the biochemical picture of low oestradiol without elevation of gonadotrophins.

challenge is to differentiate gonadotrophin deficiency from artefact or from normal outliers, the effects of ageing, obesity and insulin resistance, or other illness.

One of the most common reasons for this pattern is blood sampling in the afternoon or evening, since testosterone concentrations show a significant circadian rhythm. The first approach is therefore to repeat the sample at 09.00h, together with measurement of PRL (since hyperprolactinaemia may suppress LH/FSH and cause this picture). If a low testosterone is confirmed, then it is essential to assess remaining basal pituitary function (see above); measurement of SHBG and calculation of bioavailable testosterone may also assist, although total testosterone seems to be an equally effective measure in all but borderline cases (see Morris et al. in Further reading, below). Magnetic resonance imaging of the pituitary is also usually indicated to exclude a structural lesion.

However, even after such assessments, one is frequently left with a low or borderline testosterone and no other biochemical or radiological evidence of pituitary disease. In this situation, controversy remains: testosterone falls slowly with ageing and some workers (and many pharmaceutical companies) have advocated testosterone replacement for men with borderline testosterone and a wide variety of specific and non-specific symptoms of ill health, while others argue that this represents the normal pattern of ageing and does not require treatment. Long-term outcome studies are still awaited to answer this question. Low testosterone concentrations have also recently been described in patients with cardiac disease, with improvements in general health and cardiac function on replacement, but these data remain preliminary.

Delayed puberty

Prepubertal children have low or undetectable concentrations of testosterone or oestradiol, with low-normal concentrations of gonadotrophins. In the absence of a demonstrable pituitary lesion or other pituitary deficiencies it is, therefore, impossible to differentiate biochemically isolated gonadotrophin deficiency presenting with failure of sexual development from constitutional delayed puberty, the only difference being that the former does not resolve with time.

Constitutional delay of growth and puberty (see also Chapter 21) presents a combined problem of short stature and delayed puberty mimicking combined GH and gonadotrophin deficiency. Such patients, usually male, present in their mid-teens growing poorly, with height below the third centile and with minimal or no signs of pubertal development. The aetiology is unknown, but a family history is common and most patients eventually enter a delayed puberty with a growth spurt to attain normal adult height. Growth hormone reserve, assessed by standard dynamic tests, is usually normal. Treatment, if any, is usually with low doses of appropriate gonadal steroids to slowly advance puberty and initiate a normal growth spurt.

Hypothalamic amenorrhoea

As illustrated in [Figure 18.7](#), a variety of related conditions in the female can lead to amenorrhoea associated with the typical basal biochemistry of gonadotrophin deficiency.

Weight-related amenorrhoea frequently occurs when body weight falls below the normal healthy range for height (BMI 18.5–25 kg/m²). Anorexia nervosa is an extreme example, almost always associated with amenorrhoea and regression of secondary sexual characteristics, but dieting to achieve a weight which is ‘socially acceptably skinny’ can also induce amenorrhoea. Weight gain to a normal weight will usually restore menstruation, but chances of resumption of regular cycles are <10% with a BMI <18 kg/m² and only 50% at a low-normal BMI of 20 kg/m². Low body weight is also an important cause of delayed puberty in both sexes.

A large proportion of international standard competitive athletes develop exercise-related amenorrhoea owing to their intensive training regimens. Non-competitive athletes undertaking excessive exercise may also develop amenorrhoea or delayed puberty, often exacerbated by low body weight. The incidence of amenorrhoea is directly related to the amount of exercise taken, and if exercise is discontinued for any reason, including injury, menstruation usually returns.

Psychological stress is frequently quoted anecdotally as a cause of amenorrhoea, for example in female students attending university for the first time or undertaking examinations. ‘Hypothalamic amenorrhoea’ is a descriptive label applied to such patients and to other cases of amenorrhoea with low oestradiol and gonadotrophins and no identifiable cause, in whom organic gonadotrophin deficiency is felt to be unlikely.

There is no absolute way to differentiate these causes of hypothalamic amenorrhoea from gonadotrophin deficiency, except by their reversibility. If the disorder is mild (e.g. borderline underweight), then a clomifene test (see above) may clinch the diagnosis by inducing a rise in gonadotrophins and subsequent menstruation, but lack of response to clomifene may also occur if the disorder is severe.

If the cause of the amenorrhoea cannot be reversed, then treatment is usually with oestrogen replacement therapy to prevent the long-term effects of oestrogen deficiency.

Other isolated anterior pituitary deficiencies

Isolated ACTH deficiency occurs rarely and can only be diagnosed by full pituitary function testing. The adrenals themselves may be shown to be normal either by the finding of a low plasma ACTH prior to replacement therapy or by a depot (long) tetracosactide test. A proportion of such patients show an ACTH and cortisol response to the CRH test, indicating that the deficit is one of hypothalamic CRH secretion, but this is largely of academic interest.

Isolated TSH deficiency is also a rare occurrence.

Diabetes insipidus

Since vasopressin is synthesized by cell bodies within the hypothalamus and not controlled by the portal venous system, vasopressin deficiency is rarely found in primary pituitary disease, unless treated surgically. Diabetes insipidus is therefore a most unusual presenting symptom of a pituitary adenoma and its presence should suggest that the lesion is either a craniopharyngioma or other rarer type of tumour or inflammatory disease.

Isolated central DI may occur spontaneously. Its aetiology is unknown; some cases may be due to hypothalamic autoimmunity and it may also be familial, presenting in early childhood and typically associated with mutations in the vasopressin-neurophysin-II gene. Diabetes insipidus is discussed further in Chapter 4.

ADRENAL DISEASE

Clinical features of Addison disease

Addison disease implies primary adrenal failure: it is rare, with an incidence of 3–4/million per year and overall prevalence of 40–60/million. The disease involves destruction of the whole adrenals, today most commonly caused by autoimmune adrenal failure; in Addison’s time, tuberculosis of the adrenals was also a common cause and this remains an occasional cause today, as do other infiltrations and tumours involving the adrenals.

In the absence of another acute stressful illness, the symptoms of Addison disease are often vague and non-specific, including tiredness, weight loss, anorexia, depression, abdominal pains, nausea and vomiting and other aches and pains. The rise in plasma ACTH caused by loss of feedback results in increasing pigmentation, classically in palmar creases, scars, pressure areas and buccal mucosa, but which often simply appears as an increased ‘tan’ and is ignored. Patients who are not acutely unwell may often have normal biochemistry unless plasma cortisol concentration is measured.

During acute intercurrent illness including infection, trauma or surgery but especially with vomiting and diarrhoea or other causes of fluid loss, the undiagnosed or untreated patient may become very unwell. Features of an ‘Addisonian crisis’ include dehydration, hypotension, hypoglycaemia and a typical biochemical picture of hyponatraemia, hyperkalaemia and slightly raised plasma urea and creatinine concentrations; untreated this may

proceed to circulatory collapse and death. Treatment is a medical emergency and involves intravenous steroids (hydrocortisone 100 mg i.v. stat, repeated six-hourly), 0.9% saline and 5% dextrose if hypoglycaemic. Ideally, plasma cortisol concentration should be measured on the sample taken prior to steroid treatment, and this alone may be diagnostic if found to be low in the face of severe illness.

Patients commonly have a personal or family history of other autoimmune disease including vitiligo, primary hypothyroidism, hypoparathyroidism, premature ovarian failure, pernicious anaemia and occasionally inflammatory bowel disease. Full blood count, calcium and thyroid function should therefore always be measured both acutely and after steroid replacement (since untreated hypoadrenalism is also associated with hypercalcaemia, elevation of TSH with low free T₄, and macrocytosis, all of which may resolve when appropriate steroid replacement is given).

Congenital adrenal hyperplasia

A gene mutation resulting in deficiency of an enzyme involved in steroid biosynthesis (see Fig. 18.4) will cause inability to secrete adequate circulating cortisol, increased ACTH concentrations and adrenal stimulation due to loss of feedback, and therefore increase in steroid precursors prior to the malfunctioning enzyme. These conditions are described as congenital adrenal hyperplasia (since the adrenal increases in size under ACTH stimulation) and are summarized in Table 18.1. Of the subtypes, only 21-hydroxylase deficiency is at all common.

The diagnosis is typically made in the neonatal period in the context of ambiguous genitalia in females, or evidence of acute hypoadrenalism (including a neonatal 'salt-losing crisis'), but is sometimes delayed and, rarely, occurs at or after puberty in so-called 'late onset' 21-hydroxylase deficiency. Biochemical diagnosis involves demonstration of elevated concentrations of steroid precursors in plasma (17 α -hydroxyprogesterone in the case of 21-hydroxylase deficiency) or in urine by gas chromatography-mass spectrometry (GC-MS). In common forms of the condition, adrenal androgen concentrations are also raised (dehydroepiandrosterone sulphate (DHEAS), androstenedione, testosterone), which causes virilization in the female, but the rarer 17 α -hydroxylase deficiency prevents androgen as well as cortisol synthesis, and causes a female phenotype in genetic males as well as hypertension due to increased mineralocorticoid production.

Treatment is by steroid replacement. A variety of regimens (hydrocortisone, prednisolone and dexamethasone) is used with no hard evidence regarding the best; fludrocortisone replacement is also required in salt-losing types. In 21-hydroxylase deficiency, treatment may be monitored biochemically by measurement of 17 α -hydroxyprogesterone (a low concentration indicating overtreatment) and adrenal androgens (a high concentration indicating undertreatment). In childhood, treatment success is assessed by the close clinical monitoring of growth and puberty in the setting of a specialist paediatric endocrine service.

Assessment of adrenal incidentaloma

The increasing use of CT and MRI scanning of the abdomen in clinical medicine has revealed that small lesions consistent with adenomas in the adrenal are very common. Such incidentally discovered lesions are often described as 'adrenal incidentalomas' and require appropriate investigation to exclude malignancy and excess hormone secretion. The risk of malignancy can be assessed on radiological grounds: evidence of invasion of surrounding tissues, calcification and heterogeneity, and the lipid content of the lesion, which, if high, increases the likelihood of the adenoma being benign. However, the risk of malignancy can also be assessed simply on the basis of size, with most authorities advising removal of lesions >4 cm in diameter. In the absence of obvious clinical features of Cushing syndrome, however, the exclusion of a functioning lesion is primarily biochemical and includes exclusion of pheochromocytoma (by measurement of urine catecholamines) and Conn syndrome (see p. 760), and usually exclusion of subclinical excess cortisol secretion by a formal 48 h low-dose dexamethasone suppression test. Most clinicians would advise excision of a functioning lesion, whereas non-functioning lesions <4 cm in diameter are typically managed conservatively. There is current controversy regarding the significance of 'subclinical Cushing syndrome', which describes incomplete dexamethasone suppression in the absence of frank cushingoid symptoms. Some groups believe that this confers increased cardiovascular risk, while others believe it is a statistical 'fall out' of normal individuals.

MONITORING PITUITARY AND ADRENAL REPLACEMENT THERAPY

The clinical biochemist is often called upon to monitor the adequacy of hormone replacement therapy in pituitary disease. Clinical assessment of adequacy of replacement is imprecise and may, for example, lead to over-replacement with adrenal or thyroid hormones in response to unrelated, vague malaise.

There is no general agreement regarding the most appropriate way to monitor glucocorticoid replacement and it is difficult to validate whatever approach is taken. The authors' practice for a patient taking hydrocortisone (which is cortisol) is to measure plasma cortisol at 09.00 h (having taken the morning dose on rising), at 12.30 h (just before any lunchtime dose) and at 17.30 h (before the evening dose) and to perform a 24 h UFC measurement beginning and ending with the morning replacement dose. The aim of therapy is then to achieve a 24 h UFC within the normal range and 09.00 h cortisol within the 09.00 h normal range (to avoid over-replacement), with neither 12.30 h or 17.30 h cortisol <50 nmol/L (and ideally not <100 nmol/L to avoid under-replacement). Using these criteria, most patients require thrice-daily hydrocortisone dosage, typically 10 mg on rising, 5–10 mg at lunchtime and 5 mg in the early evening, which is consistent with current estimates of normal cortisol secretion rates. Random measurements of plasma cortisol on replacement therapy are of little or no value.

In primary adrenal failure, mineralocorticoid replacement is also required, and adequacy of replacement can be assessed biochemically by measuring lying plasma renin (which is markedly elevated in untreated patients and ideally should be normal on fludrocortisone replacement). Since the renin–angiotensin system is intact in patients with pituitary disease, fludrocortisone replacement, with monitoring of plasma renin, is not necessary.

The aim of levothyroxine replacement in hypopituitarism should be to achieve a plasma free T4 in the middle- or upper-third of the normal range. Measurements of TSH are clearly unhelpful in this context, and a common mistake is to reduce the dose of thyroxine because of the incorrect conclusion that a low TSH indicates over-replacement.

Biochemical assessment of gonadal steroid replacement presents greater difficulty. In a male on testosterone ester depot preparations, plasma testosterone may reasonably be measured a few days after an injection and just prior to the next dose to ensure that adequate concentrations are being achieved. Plasma testosterone measurement is also useful in patients on transdermal or mucoadhesive testosterone preparations. In patients on oral testosterone undecanoate therapy, however, there is preferential conversion of testosterone to dihydrotestosterone, and plasma total testosterone concentrations are frequently below normal despite adequate replacement therapy. In the female, the best approach to therapy is probably to use the endometrium as a natural *in vivo* bioassay for oestrogen and aim simply to achieve regular withdrawal bleeds. It is possible to measure plasma oestradiol in patients receiving oral or transdermal oestradiol replacement, but other oral preparations of modified synthetic or equine oestrogens cannot logically be monitored biochemically.

Desmopressin replacement therapy can be monitored by measurement of plasma osmolality, aiming to maintain a normal osmolality and avoid dehydration or fluid overload. Patients should be instructed to allow the diuresis to resume prior to each treatment dose.

Patients with hypopituitarism and with primary adrenal failure require appropriate information and education about the nature of their replacement therapy and, in particular, the need to increase steroid hormone replacement in the face of intercurrent illness. They should be advised to carry a 'Steroid card' or alert bracelet and ensure that all doctors and dentists who carry out treatment are aware of their condition and the need for increased treatment during stress.

CONCLUSION

In the last decades of the 20th century, clinical neuroendocrinology was revolutionized by the introduction of reliable assays for the relevant pituitary and adrenal hormones. The relationship between endocrinologist and clinical biochemist must remain a close one, since the biochemical investigations to be performed depend closely on the clinical assessment of the patient, while the treatment which the patient will receive is equally dependent on the biochemical results obtained. In this chapter, it is hoped that a logical framework for such a harmonious relationship has been laid down.

Further reading

- Arnaldi G, Angeli A, Atkinson AB et al. Diagnosis and complications of Cushing's syndrome: a consensus statement. *J Clin Endocrinol Metab* 2003;88:5593–602.
- Baylis PH. Diabetes insipidus. *J R Coll Physicians Lond* 1998;32:108–11.
- Biller BM, Grossman AB, Stewart PM et al. Treatment of adrenocorticotropin-dependent Cushing's syndrome: a consensus statement. *J Clin Endocrinol Metab* 2008;93:2454–62.
- Clark PM, Neylon I, Raggatt PR et al. Defining the normal cortisol response to the short Synacthen test: implications for the investigation of hypothalamic-pituitary disorders. *Clin Endocrinol (Oxf)* 1998;49:287–92.
- Dickstein G. Hypothalamo-pituitary-adrenal axis testing: nothing is sacred and caution in interpretation is needed. *Clin Endocrinol (Oxf)* 2001;54:15–6.
- Giustina A, Chanson P, Bronstein MD et al. A consensus on criteria for cure of acromegaly. Acromegaly Consensus Group. *J Clin Endocrinol Metab* 2010;95:3141–8.
- Lindholm J. The insulin hypoglycaemia test for the assessment of the hypothalamic-pituitary-adrenal function. *Clin Endocrinol (Oxf)* 2001;54:282–6.
- Melmed S, Polonsky KS, Reed Larsen P et al. editors. *Williams' textbook of endocrinology*. 12th ed Philadelphia: Elsevier Saunders; 2011.
- Morris PD, Malkin CJ, Channer KS et al. A mathematical comparison of techniques to predict biologically available testosterone in a cohort of 1072 men. *Eur J Endocrinol* 2004;151:241–9.
- National Institute for Health and Care Excellence. Growth hormone deficiency (adults) – human growth hormone (Technology Appraisal No. 64): <http://www.nice.org.uk/page.aspx?o=83406>; [Accessed October 2013].
- Nieman LK, Biller BM, Findling JW. The diagnosis of Cushing's syndrome: an endocrine society clinical practice guideline. *J Clin Endocrinol Metab* 2008;93:1526–40.
- Pekic S, Popovic V. Diagnosis of growth hormone deficiency in adults. In: Ho K, editor. *Growth hormone related diseases and therapy. A molecular and physiological perspective for the clinician*. New York: Humana Press; 2011.

APPENDIX 18.1 TEST PROTOCOLS

Assessment of basal pituitary function

A basal 09.00 h blood sample for measurement of:

- cortisol
- free T4
- TSH
- testosterone (male) or oestradiol (female)
- LH, FSH
- prolactin
- IGF-1 in the context of proven pituitary disease (but not simply as a screening test)
- osmolality^a.

^aSimultaneous random urine for osmolality.

(^aAssessment of posterior pituitary function is only required if indicated by the clinical context.)

Insulin stress test

Indications: assessment of ACTH/cortisol and GH reserve. *Contraindications:* ischaemic heart disease, epilepsy or unexplained blackouts, severe longstanding hypoadrenalism.

Precautions: electrocardiograph must be normal, basal 09.00 h cortisol must be >100 nmol/L. Free T4 must be normal.

Procedure: fasting from midnight, *i.v.* cannula, soluble insulin, 0.15 U/kg *i.v.* at 09.00 h. Observe and record signs of hypoglycaemia. Hypoglycaemia only needs to be

reversed with i.v. dextrose if severe and prolonged or with impending or actual loss of consciousness or fits.

Sampling: blood for glucose, cortisol, GH at 0, 30, 45, 60, 90 and 120 min.

Interpretation: blood glucose must fall below 2.2 mmol/L to achieve an adequate hypoglycaemic stress. If adequate hypoglycaemia is not achieved then cortisol or GH deficiency cannot be diagnosed. Untreated hypothyroidism can also give subnormal results.

A normal response is traditionally defined as a rise in plasma cortisol to a peak of ≥ 550 nmol/L (but see comments in text (p. 355) on normal ranges) and a rise in plasma GH to ≥ 7 μ g/L (most normal individuals actually rise to a peak >15 μ g/L). 'Severe GH deficiency' is arbitrarily defined as a GH peak <3 μ g/L.

Patients who show such a normal cortisol response can withstand major surgery without corticosteroid replacement. Patients with responses in the range 400–550 nmol/L may not require regular replacement therapy, but do require cover for major illnesses and surgery. All other patients with subnormal responses require hydrocortisone replacement.

Children with bone age <12 years and GH peak <6.7 μ g/L generally benefit from GH therapy.

Short tetracosactide (synacthen) test

Indication: screening test for adrenocortical insufficiency, including ACTH deficiency.

Procedure: synacthen (tetracosactide) 250 μ g i.v. or i.m. Plasma cortisol before and after 30 min.

Interpretation: plasma cortisol of ≥ 600 nmol/L at 30 min is highly correlated with a normal ACTH response to hypoglycaemia and a normal pituitary-adrenal axis. However, normal responses vary with gender and between commonly used cortisol assays – an appropriate normal range for the assay methodology must, therefore, be used (see p. 355).

Responses >600 nmol/L can be considered normal for almost all assays and circumstances. Peak cortisol responses between 400 and 600 nmol/L may be regarded as equivocal – they require appropriate patient education and may need confirmation by a formal IST (if not contraindicated).

Water deprivation test

Indications: diagnosis of vasopressin deficiency.

Procedure: if mild DI is suspected, the patient should have no fluids from midnight the night before (dry foods allowed). If severe DI is likely, free fluids and a light breakfast should be allowed prior to the test.

Commence test at 09.00h or earlier if practicable. Measure plasma and urine osmolality basally, then hourly until the end of the test (after 8 h or when diagnostic osmolalities achieved). Record all urine volumes passed.

Osmolalities should be measured prospectively as the test proceeds. If this is not possible, weigh the patient every 2 h and arrange for urgent plasma osmolality if body weight falls by more than 3%.

When osmolalities diagnostic of DI have been achieved, give desmopressin 2 μ g i.m., allow free fluids and measure urine osmolality hourly for 2 h.

Interpretation: diabetes insipidus is diagnosed by a rise in plasma osmolality above 300 mmol/kg, without appropriate concentration of urine osmolality (to >600 mmol/kg). Adequate concentration of the urine after desmopressin indicates central, cranial DI and lack of concentration indicates nephrogenic DI.

Normal vasopressin secretion is shown by concentration of the urine to >600 mmol/kg, in the presence of a normal plasma osmolality (280–295 mmol/kg).

Normal concentration of the urine at the expense of elevation of plasma osmolality (>295 mmol/kg) is not a strictly normal response and may indicate mild DI or other subtle defects of thirst, osmoregulation or vasopressin secretion. Further investigations may be indicated.

Failure to concentrate either plasma or urine osmolality by the end of the test is non-diagnostic, but usually represents primary polydipsia. Such patients typically start the test with very dilute urine and often dilute plasma as a result of fluid overload. If clinically indicated, more prolonged water deprivation is required to make the diagnosis.

Glucose tolerance test for the diagnosis of acromegaly

Procedure: fasting from midnight, i.v. cannula, 75 g glucose at time 0.

Sampling: blood glucose and plasma GH at 0, 30, 60, 90, 120 and 150 min.

Normal response: growth hormone suppresses to <1 μ g/L after glucose in normal subjects.

Interpretation: failure of growth hormone to suppress completely after glucose is highly suggestive of acromegaly, even if a substantial fall has occurred.

CRH test

Indications: differential diagnosis of Cushing syndrome.

Differential diagnosis of isolated ACTH deficiency.

Procedure: i.v. cannula, human CRF-41, 100 μ g i.v. at 09.00h.

Notes: in North America it is usual to administer CRH in the evening rather than early morning. Ovine CRF-41 was originally used but is not readily available in the UK. Most patients notice facial flushing after CRH, but there are no other side-effects.

Sampling: plasma cortisol and ACTH at –15, 0, 15, 30, 45, 60, 90 and 120 min.

Normal response: plasma cortisol rises by 25% or more above the basal concentration. The peak response to hCRF-41 in normal individuals is <600 nmol/L (<800 nmol/L with ovine CRF-41).

Interpretation: patients with Cushing disease show a normal or exaggerated response to CRH. Patients with ectopic ACTH-secreting tumours and adrenal tumours show no response (see text).

Responses in ACTH deficiency remain to be defined but a response of plasma ACTH would imply hypothalamic CRH deficiency. Patients with depression show a normal cortisol but reduced ACTH response. Patients with massive obesity show a blunted cortisol response.

Thyroid dysfunction

Colin M. Dayan • Onyebuchi E. Okosieme • Peter Taylor

CHAPTER OUTLINE

INTRODUCTION 373

NORMAL THYROID PHYSIOLOGY 374

- The thyroid gland 374
- Biological actions of thyroid hormones 374
- Synthesis, storage and release of thyroid hormones 375
- Iodine and thyroid hormone synthesis 376
- Transport of thyroid hormones in blood 376
- Free hormone hypothesis 376
- Entry of thyroid hormone into tissues 377
- Thyroid hormone deiodination and regulation of extrathyroidal T3 production 377
- Catabolism of thyroid hormones 377
- Nuclear action of thyroid hormones 377
- Control of thyroid hormone synthesis and secretion 378
- Extrathyroidal factors that may affect thyroid function 379

THE EVALUATION OF THYROID FUNCTION 382

- Clinical evaluation of thyroid status 382
- In vitro tests of thyroid activity and pituitary–thyroid status 383
- Measurement of thyroid stimulating hormone 383
- Free T4 and free T3 measurements 383
- Methods for measuring free thyroid hormones 383
- Validity of commercial methods for free hormone analysis 383
- Nomenclature of free thyroid hormone assays 384
- Total T4 and total T3 384
- Selective use of thyroid function tests 384

- Interpreting results of thyroid function tests 385
- Common situations in which TSH results may be misleading 385
- Reference ranges and significant changes 385
- Miscellaneous tests 386
- Autoantibodies to thyroidal antigens 386
- Imaging the thyroid 387

HYPERTHYROIDISM 388

- Clinical features 388
- Causes of hyperthyroidism 389
- Hyperthyroidism or non-thyroidal illness? 394

HYPOTHYROIDISM 394

- Clinical features 394
- Causes of hypothyroidism 396
- Treatment of hypothyroidism 397

THYROIDITIS 398

- Thyroiditis producing hyperthyroidism 398
- Hypothyroidism resulting from Hashimoto thyroiditis 399
- Other forms of thyroiditis 399
- Hypothyroidism and the postpartum period 399

NEOPLASIA 399

- Diagnosis 399
- Treatment 400
- Tumour markers 400

SYNDROMES OF RESISTANCE TO THYROID HORMONES 400

SCREENING 401

INTRODUCTION

Thyroid hormones are essential for normal growth, development and metabolism, and their production is tightly regulated through the hypothalamic–pituitary–thyroid axis. Thyroid disease is common, particularly in women, with a prevalence in the community of 3–5%. With the exception of iodine deficiency, which affects millions of people worldwide, diseases that directly

affect the thyroid gland are the most common causes of thyroid dysfunction. Pituitary disease and the use of certain drugs that alter thyroid hormone synthesis or metabolism can also give rise to thyroid dysfunction. Any severe illness can produce abnormalities in the results of thyroid function tests that resolve as the patient's illness improves. Once diagnosed, thyroid disease is usually easily treated, with an excellent long-term outcome for most patients. This chapter outlines thyroid physiology and the

pathways of thyroid hormone synthesis and metabolism. The investigations used in diagnosis and management of thyroid disorders are described, together with guidance on their interpretation. The various thyroid disorders are described together with current views on their treatment.

NORMAL THYROID PHYSIOLOGY

The thyroid gland

The thyroid gland is brownish-red in colour and consists of left and right lobes connected by a midline isthmus, typically forming an 'H' or 'U'. It is located anteriorly in the neck deep to the platysma, sternothyroid and sternohyoid muscles. The isthmus lies just below the cricoid cartilage and the lobes extend upward over the lower half of the thyroid cartilage. The pretracheal fascia encloses the thyroid gland and attaches it to the trachea and larynx. This explains why the thyroid gland is seen to move upwards with swallowing.

The thyroid is a highly vascular organ with blood flow ~ 5 mL/min/g of tissue (nearly twice that of the kidneys). It consists of thousands of follicles, each a spheroidal sac of epithelial cells (thyrocytes) surrounding a lumen containing colloid, mainly comprising thyroglobulin (Fig. 19.1). It also contains parafollicular C cells, which secrete calcitonin; calcitonin is discussed in Chapter 6.

Embryologically, the thyroid is derived from the floor of the pharyngeal cavity and migrates from the base of the tongue to its final position in the neck along the thyroglossal duct. Abnormalities of this migration give

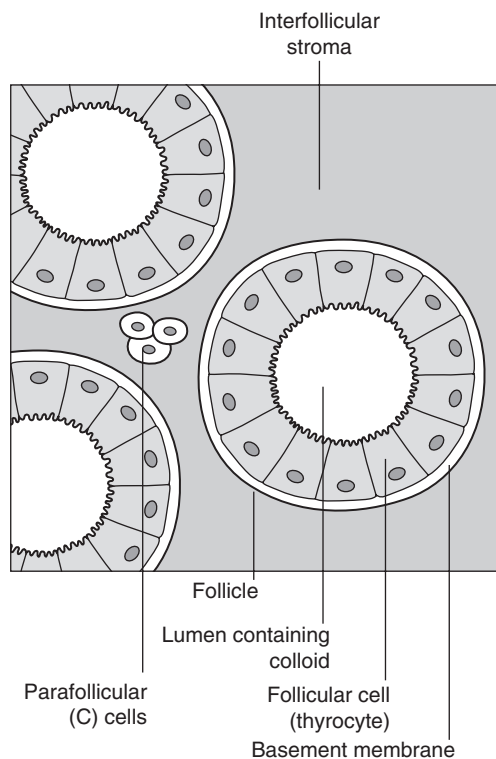


FIGURE 19.1 ■ Structure of the thyroid follicle.

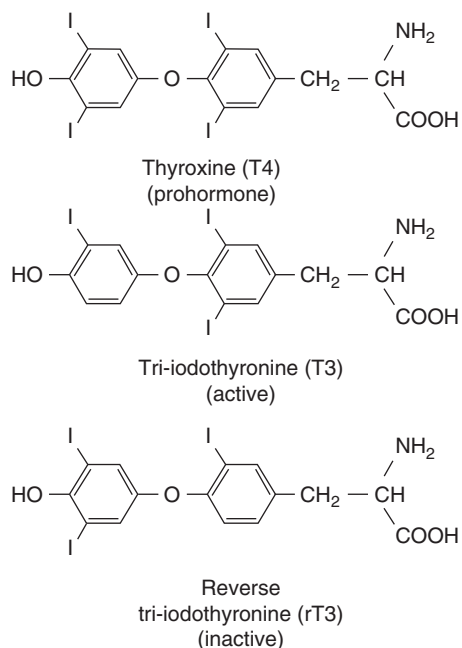


FIGURE 19.2 ■ Structures of the iodothyronines.

rise to ectopic glands that may not function normally. Occasionally, the gland is completely absent.

The primary function of the thyroid is to synthesize and secrete thyroid hormones (Fig. 19.2). Under normal circumstances the gland synthesizes approximately $80 \mu\text{g}$ (110 nmol) of thyroxine (T4) per day and $6 \mu\text{g}$ (10 nmol) of 3,5,3'-tri-iodothyronine (T3).

Biological actions of thyroid hormones

Thyroid hormones produce enhancement of many intracellular events and promote differentiation and growth; they are essential for normal fetal and neonatal development.

Thyroid hormones increase mitochondrial oxidative phosphorylation and maintain amino acid and electrolyte transport into cells. They increase calorogenesis and oxygen consumption in most tissues, though not in brain. Thyroid hormones stimulate the synthesis of proteins, including structural proteins and enzymes. They regulate all aspects of carbohydrate metabolism, including increasing gluconeogenesis and accelerating insulin degradation. Thyrotoxicosis can, therefore, lead to a deterioration of glycaemic control in diabetes mellitus.

Stimulation of lipid metabolism leads to a fall in plasma cholesterol concentration, as degradation is increased to a greater extent than synthesis. Bone turnover is stimulated, resorption more so than mineralization, raising plasma calcium concentration in some patients. In general, the effect of thyroid hormones on the turnover of any substance will depend on the relative effects on synthesis and degradation. The actions of thyroid hormones increase demand for coenzymes and vitamins.

Complex interactions occur with the sympathetic nervous system and many of the signs and symptoms of hyperthyroidism, including palpitations, tremor and

sweating may be attributed to these. For example, thyroid hormones amplify the adrenergic system through the α_1 and β_3 adrenergic receptors and are also the main determinant of basal metabolic rate (BMR). The increase in metabolic rate involves inhibition of AMP-activated protein kinase (AMPK) in the hypothalamus. However, the mechanisms responsible for many interactions between thyroid hormone and the sympathetic nervous system have not been identified.

Most changes in biochemical parameters produced by changes in thyroid status are not sufficiently specific or of such magnitude to be useful as tests of thyroid function, though the association may be sufficiently close that the finding of an abnormality should provoke a request for testing for previously unsuspected thyroid disease. For example, hypothyroidism is a relatively common and easily treatable cause of raised plasma cholesterol; however, the effect on lipids in subclinical hypothyroidism is modest. Equally, it may be helpful (and avoid unnecessary investigations) to anticipate these changes in cases of known thyroid dysfunction and to expect their normalization after treatment of the thyroid disorder. Some examples are listed in [Box 19.1](#).

Synthesis, storage and release of thyroid hormones

Synthesis of T₄ and T₃ occurs on thyroglobulin (Tg), a glycoprotein of molecular weight 660 000 Da that contains many tyrosyl residues. Thyroglobulin is synthesized by the thyrocytes and exported to be stored within the colloid of the follicular lumen. Incorporation of iodide into Tg requires hydrogen peroxide and thyroid peroxidase (TPO), an enzyme that is synthesized within the follicular cell and transported to the apical membrane. Thyroid hormones are synthesized at the interface between the

BOX 19.1 Effects of thyroid hormones on metabolic indices

Parameters increased by hyperthyroidism

- Basal metabolic rate
- Plasma
 - Calcium
 - Sex hormone-binding globulin
 - Angiotensin-converting-enzyme
 - Liver enzymes (alkaline phosphatase, aminotransferases, γ -glutamyl transferase)
- Glucose tolerance tests may show a degree of glucose intolerance

Parameters increased by hypothyroidism

- Plasma
 - Cholesterol
 - Creatine kinase (CK-MM isoenzyme)
 - Creatinine
 - Thyroxine binding globulin
 - Prolactin

Parameters decreased by hypothyroidism

- Plasma sodium

apical membrane of the thyrocyte and the colloid of the follicular lumen.

The process of thyroid hormone synthesis, storage and secretion requires a series of highly regulated steps ([Fig. 19.3](#)).

- *Trapping of iodide.* Iodide from plasma is actively transported by a sodium-iodine symporter situated in the basal membrane of the thyrocyte. This process allows iodide to be actively taken up against a steep concentration gradient. At the apical membrane of the cell, pendrin, a relatively non-specific anion transporter, mediates iodide efflux. The

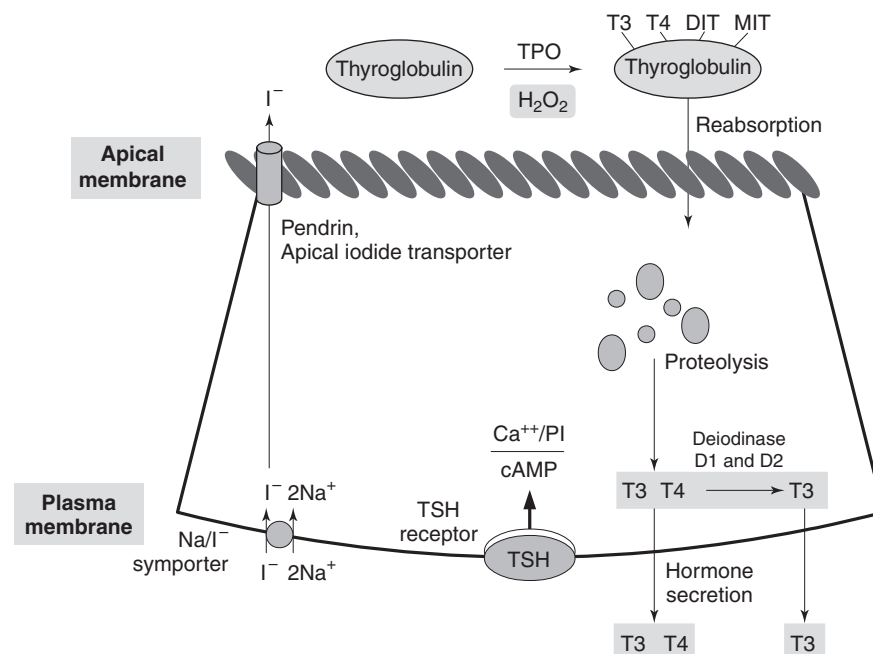


FIGURE 19.3 ■ Sites of biochemical processes within the follicular cell (thyrocyte).

sodium-iodide symporter is competitively inhibited by anions of similar size to iodide. Perchnetate is also transportable and thus is used for radioactive imaging of the gland. Perchlorate is used to block the uptake of iodide, for example after accidental exposure to radioactive iodide. Thiocyanates (found in certain foods, e.g. cassava) competitively inhibit the iodide pump but are not taken up into the gland.

- *Oxidation of iodide to iodine by thyroid peroxidase.* This occurs on the luminal side of the apical membrane and requires TPO and hydrogen peroxide, which is generated by a calcium-dependent flavoprotein enzyme system situated at the apical membrane.

Recently, the dual oxidase molecules DUOX1 and DUOX2, which contain peroxidase-like and NADPH oxidase-like domains have been identified as essential for H_2O_2 production. They require maturation or activation factors (DUOXA1 or 2) for translocation of DUOX from the endoplasmic reticulum to the apical plasma membrane where H_2O_2 production occurs. Mutations in DUOX2 and DUOXA2 have both recently been identified as causes of hypothyroidism, resulting from diminished H_2O_2 production.

Halogenases collect the iodide removed from thyroglobulin as it is broken down in the cell.

Antithyroid drugs such as propylthiouracil and mercaptoimidazoles (methimazole and carbimazole) exert their action through inhibiting the action of TPO.

- *Incorporation of iodine into tyrosyl residues on thyroglobulin.* Mono-iodotyrosine and di-iodotyrosine (MIT and DIT) are formed through the actions of TPO (organification).
- *Coupling of two iodotyrosyl residues in the thyroglobulin molecule.* This is also catalysed by TPO and produces T3 and T4 that remain linked to Tg. When iodine supply is limited, the proportion of T3 produced on Tg increases. Iodinated Tg is stored in the follicular lumen.
- *Internalization of Tg and release of T4 and T3.* When there is a demand for thyroid hormone, this is signalled by an increase in plasma thyroid stimulating hormone (thyrotrophin; TSH) concentration. Thyroglobulin is then internalized by pinocytosis and appears as colloid droplets that fuse with lysosomes and undergo proteolytic degradation to release T3 and T4. Any MIT and DIT is deiodinated and the iodine conserved.
- *Delivery of T4 and T3 into the circulation.* This is probably achieved by thyroid hormone transporters. The large stores of T4 and T3 incorporated into thyroglobulin allow secretion of T4 and T3 more quickly when required than if it had to be synthesized.

Thyroid stimulating hormone appears to stimulate each of the above processes. It also stimulates the expression of many genes in thyroid tissue including thyroglobulin, sodium iodide symporter (NIS) and interleukin-8. It also causes thyroid hyperplasia and hypertrophy.

Iodine and thyroid hormone synthesis

Iodine deficiency represents a major public health issue for over two billion people who live in areas of iodine deficiency and especially for the 43 million in whom the deficiency is severe enough to cause learning difficulties. The recommended intake of iodine is 150 $\mu\text{g}/\text{day}$ for adults but this may rise to 250 μg during pregnancy and lactation. Iodine is obtained from the ingestion of foods such as seafood, seaweed, kelp, dairy products, some vegetables (e.g. soybeans and potatoes) and iodized salt.

High iodine intakes can induce hypothyroidism, goitre and sometimes hyperthyroidism (see later). A large excess of iodide, when given acutely, results in acute inhibition of thyroid hormone release. It also inhibits the adenylate cyclase response to TSH (see below) and iodination of thyroglobulin; this is termed the Wolff–Chaikoff effect. After a few days of exposure to high iodide concentrations, thyroidal iodide uptake is very low, the intrathyroidal iodide concentration falls and the synthesis of iodinated thyroglobulin recommences. This effect can be used to prepare a thyrotoxic patient for thyroidectomy. However, if high iodine intake persists for more than ten days, then thyrotoxicosis can occur in individuals with a pre-existing multinodular goitre or those from iodine deficient areas; this is termed the Jod–Basedow effect.

Transport of thyroid hormones in blood

When T4 and T3 are secreted into the bloodstream, they become bound to plasma carrier proteins. Three proteins share this function: in order of decreasing affinity for iodothyronines they are thyroxine binding globulin (TBG); thyroxine binding prealbumin (TBPA, also known as transthyretin), and albumin. These binding proteins bind approximately 70% (TBG), 20% (TBPA) and 10% (albumin) of the circulating thyroid hormones. T4 is more tightly bound than T3 to each of these proteins. Approximately 0.02% of T4 and 0.2% of T3 are in the unbound or 'free' form in plasma in normal subjects, so that, although the total molar concentration of T4 is some 50 times that of T3, the concentration of free T4 is only about three times that of free T3. The half-life of T4 is 5–7 days; that of T3 is 1–2 days.

Free hormone hypothesis

For most of their actions, thyroid hormones need to enter the cell and bind to nuclear receptors. The free hormone hypothesis suggests that only the unbound (free) fraction of thyroid hormone is able to cross the plasma membrane and enter the nucleus. This hypothesis is supported by the observation that, in clinical practice, thyroid status correlates better with the concentration of the free hormone fraction than with the protein-bound fraction. The precise function of TBG and TBPA is unclear, but a number of roles have been suggested including acting as a reserve or buffer to prevent rapid changes in free hormone concentration in blood and also to prevent loss of thyroid hormones through filtration and excretion at the kidney. One further hypothesis

is that TBG may be involved in targeting the delivery of thyroid hormone in a differential manner to different organs or facilitating the efflux of thyroid hormones from some tissues. Whatever their function, the absence or overexpression of the transport proteins is not associated with any obvious detrimental effects. There are difficulties in measuring the concentrations of free hormones and this is discussed later in this chapter.

Entry of thyroid hormone into tissues

Specific transporters appear to be required to facilitate the entry of the free thyroid hormones into cells. A number of important thyroid hormone transporters have been characterized. The organic anion transporting polypeptides (OATPs) are one group of important transporters. OATP1C1 appears to be specific for T4 and reverse tri-iodothyronine (3,3′5′-tri-iodothyronine: rT3); it is expressed exclusively in the capillaries of the brain and may be important for T4 transport across the blood–brain barrier. Two L-type amino acid transporters (LAT1 and LAT2) have been identified among the members of the heterodimeric amino acid transporter family, that transport thyroid hormone in many tissues. Human monocarboxylate transporter 8 (MCT8) is another active thyroid hormone transporter. It has a preference for T3 and is expressed in a variety of tissues including the brain, where it appears to be involved in the uptake of T3 by neurons, the primary target for thyroid hormones during brain development. The MCT8 gene is located on the X chromosome and mutations in the gene have been identified in a small number of boys with severe psychomotor retardation and elevated plasma T3 concentration (Allan–Herndon–Dudley syndrome). These mutations result in an inhibition of T3 supply to neurons leading to a defect in brain development. The characteristic pattern of thyroid function tests remains unexplained.

Entry of thyroid hormone into liver cells appears to be ATP-dependent. In rat hepatocytes, it is inhibited by a range of compounds that accumulate in plasma during illness. These compounds include free fatty acids, bilirubin and indoxyl sulphate. The uptake of thyroid hormones by the pituitary appears to occur through a different mechanism from that occurring in the liver. For example, unlike in the liver, the pituitary thyroid hormone transporters are not energy dependent; they can maintain or increase the uptake of T4 and T3 in low energy states and are not inhibited by bilirubin.

Thyroid hormone deiodination and regulation of extrathyroidal T3 production

Thyroxine is produced entirely by the thyroid gland, approximately 80 µg being produced each day. However, only 20% of T3 is produced directly from the thyroid gland; 80% is produced by the extrathyroidal deiodination of T4. Approximately 6 µg of the T3 in blood originates from thyroidal synthesis; 25 µg is formed by 5-deiodination of T4 in extrathyroidal tissues, a reaction catalysed by the D1 and D2 isoenzymes of iodothyronine

deiodinase. T4 is considered to be a prohormone, T3 being responsible for the biological actions of thyroid hormones. Most tissues express one or more of a family of three iodothyronine deiodinases (D1, D2 and D3) that are responsible for providing a local source of T3 from T4 or for providing a source of T3 in plasma. In humans, it has proved difficult to determine which deiodinase and which tissues are responsible for providing most of the circulating T3. Classically it has been thought that D1 in liver and kidney provided most circulating T3, but some studies have suggested that, in humans, D2 may also provide an important source. Genetic association studies may provide more conclusive evidence. To date common genetic variants (single nucleotide polymorphisms) in *DIO1* have been robustly associated with variation in the T3/T4 ratio in serum at genome wide levels of significance ($p = < 1 \times 10^{-8}$), but as yet no associations have been observed with single nucleotide polymorphisms in *DIO2*.

Figure 19.4 summarizes the characteristics of the three deiodinases. All are selenoenzymes requiring adequate intake of dietary selenium for their expression. D1 carries out either 5′-deiodination, giving T3, or 5-deiodination, giving the inactive compound rT3. D2 appears to provide a very important local source of T3 in some tissues, including the brain, pituitary and brown adipose tissue. D3 catalyses the conversion of T4 to rT3 and is thought to be an important extrathyroidal control mechanism to regulate delivery of T3 to its receptors, particularly in the developing fetus.

Catabolism of thyroid hormones

Sulphation, glucuronidation, deamination, oxidative decarboxylation, ether cleavage and deiodination form the main routes for the inactivation and degradation of thyroid hormone. These metabolites are excreted via bile or in urine.

Nuclear action of thyroid hormones

Tri-iodothyronine exerts its effects through interaction with nuclear thyroid hormone receptors (TR) that have a high affinity and high specificity for T3. Nuclear thyroid hormone receptors are a family of ligand-regulated transcription factors that are associated with chromatin. Several α and β isoforms of TR are produced and these usually dimerize with the retinoid X receptors (RXRs). TR α 1 is widely expressed with particularly high expression in cardiac and skeletal muscles, whereas TR β 1 is predominantly expressed in the brain, liver and kidney and TR β 2 is primarily limited to the hypothalamus and the pituitary.

The TR/RXR complex binds to target response elements located in the promoter region of the target genes. The formation of the T3-TR/DNA complex and subsequent recruitment of a variety of transcriptional coactivators leads to activation of the target genes, giving increased mRNA and protein production. In some circumstances, gene expression may be switched off. Mutations in the *TR β* gene lead to a decreased affinity of the receptor for T3, leading to the autosomal dominant clinical syndrome of ‘thyroid hormone resistance’ (discussed later in this chapter).

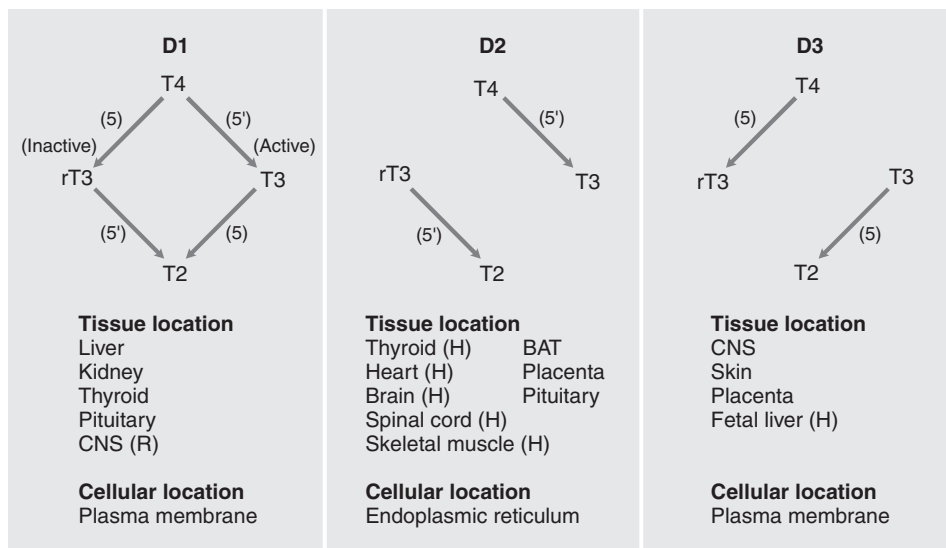


FIGURE 19.4 ■ Characteristics of the iodothyronine deiodinases. In non-thyroidal illness, 5'-mono-deiodination is impaired leading to a decreased production of T3 and impaired breakdown of rT3. BAT, brown adipose tissue; CNS, central nervous system; (H), human not rat; (R), rat not human; T₂, 3,3'-di-iodothyronine. (From Beckett G J, Arthur J R. *Journal of Endocrinology* 2005; 184:455–465. © Society for Endocrinology, with permission.)

Control of thyroid hormone synthesis and secretion

Classic feedback regulation

The most important regulator of thyroid homeostasis is TSH. This dimeric peptide hormone comprises a β -subunit, required for binding to the TSH receptor, and an α -subunit. The α -subunit is the same as that of luteinizing hormone, follicle stimulating hormone and human chorionic gonadotrophin. The β -subunit is specific for TSH: both subunits are required for bioactivity. The subunits contain carbohydrate chains that are essential for the biological activity of the molecule. Modifications to these carbohydrate residues occur in different thyroid states, which alter both the bioactivity of the molecule and its plasma half-life. For example, in primary hypothyroidism, TSH has increased bioactivity and an increased plasma retention time, while in secondary hypothyroidism the bioactivity of the molecule is diminished. Most immunoassays are unable to recognize modifications to the carbohydrate chains. There is a circadian rhythm of TSH secretion, plasma concentrations being highest between midnight and 04.00h and lowest at about midday. Thyroid stimulating hormone release also occurs in a pulsatile fashion. These changes in TSH are not reflected in the serum thyroid hormone concentrations.

The production of TSH is controlled by a stimulatory effect of the hypothalamic tripeptide, thyrotrophin releasing hormone (TRH, thyroliberin), mediated by a negative feedback from circulating free T3 and free T4 (Fig. 19.5). Free T3 and free T4 inhibit TSH synthesis and release directly, by inhibiting transcription of the TSH subunit genes and indirectly by inhibiting TRH release. Free T3 and free T4 also decrease TSH glycosylation and therefore its biological activity. It is thought

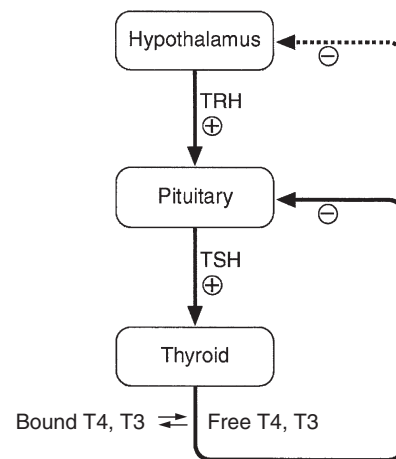


FIGURE 19.5 ■ The 'classic' hypothalamo-pituitary-thyroid axis. The production of thyroid stimulating hormone (TSH) is controlled by a stimulatory effect of the hypothalamic tripeptide, thyrotrophin releasing hormone (TRH, thyroliberin), mediated by a negative feedback from circulating free T3 and free T4. Thyroid stimulating hormone may also exert autoregulatory fine control on its own production. Dopamine, somatostatin, glucocorticoids, leptin and catecholamines all may have an influence on the axis.

that the hypothalamus, via TRH, sets the level of thyroid hormone production required physiologically, and that the pituitary acts as a 'thyroid-stat' to maintain the level of thyroid hormone production that has been determined by the hypothalamus. Twin studies have shown that in individuals without thyroid disease, TSH and thyroid hormone concentrations are largely genetically determined. The setpoint in an individual is relatively narrow with intraindividual variation being less than half the interindividual variation.

Other mechanisms

The classic feedback regulation of TSH release described above is an oversimplification. Dopamine, somatostatin and glucocorticoids inhibit TSH release, and these agents, together with cytokines, may be important modifiers of plasma TSH, particularly in patients with non-thyroidal illness (NTI) (Fig. 19.6). Leptin and catecholamines regulate TSH indirectly by stimulating TRH production. There is accumulating evidence that TSH can provide autoregulation via a direct action of TSH on the hypothalamus and at the pituitary. In the pituitary, folliculo-stellate cells have been identified that express TSH receptors: interaction of TSH with these cells may produce fine control on TSH release through the release of cytokines, which in turn inhibit TSH release from thyrotrophs. It has been argued that interaction of TSH receptor antibodies with the folliculo-stellate cells explains why patients with Graves disease may continue to have suppressed TSH weeks or months after normal thyroid hormone concentrations have been achieved through carbimazole therapy.

Thyroid stimulating hormone modifies thyroid hormone synthesis by binding to a specific receptor on the surface of the thyrocyte. The TSH receptor is a single protein with a large extracellular amino-terminal domain involved in the binding of TSH, seven transmembrane domains and a short intracellular carboxy-terminal domain involved in the activation of G-protein modulators of the adenylate cyclase-protein kinase-A system. These characteristics are shared with receptors for gonadotrophins: their extracellular domains show some 40% homology, the remainder being more highly conserved. This may explain the weak thyroid-stimulating activity of human chorionic gonadotrophin (hCG). Binding of TSH results in activation of adenylate cyclase and accumulation of cyclic AMP. The calcium and phosphoinositol signalling pathways may also be activated by TSH. After about an hour, an increase in release of thyroid hormones is seen. In the longer term, TSH increases the synthesis of iodinated thyroglobulin and also causes a general increase in the metabolism, size and activity of the follicular cells.

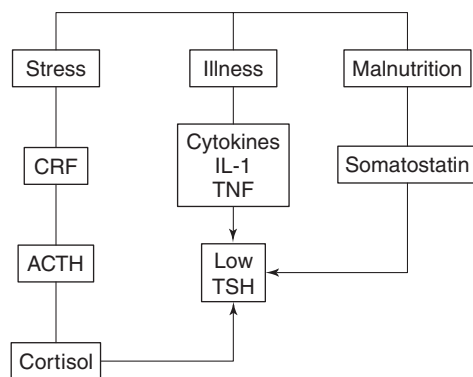


FIGURE 19.6 ■ The hypothalamo-pituitary-thyroid axis in non-thyroidal illness. Dopamine, somatostatin, and glucocorticoids inhibit TSH release, and these agents together with cytokines may be important modifiers of serum TSH, in particular in patients with non-thyroidal illness. ACTH, adrenocorticotrophic hormone; CRF, corticotrophin releasing factor; IL-1, interleukin-1; TNF, tumour necrosis factor.

Extrathyroidal factors that may affect thyroid function

Age

Fetus. Thyroxine and TSH are detectable in fetal plasma at 10–12 weeks of gestation. Plasma concentrations of total and free T₄, total and free T₃ and TBG increase during gestation, total and free T₄ reaching adult concentrations at about 36 weeks, but total and free T₃ reaching only the lower limit of the adult range at this time. The concentration of TSH also increases with gestational age but is within or above the adult reference range throughout gestation. Marked time- and organ-specific changes occur throughout gestation in the expression of the deiodinases, which are thought to coordinate the orderly maturation of enzyme systems responsible for development of the brain and the other organ systems in the fetus.

Neonate. During the first 24 h after birth, there is a rapid, transient increase in the release of TSH, T₃ and T₄ that is considered to be an adaptive response to birth. Thyroid stimulating hormone peaks during the first 30 min, which then stimulates T₃ and T₄ production during the first 24–36 h of postnatal life. This effect may be attenuated in infants delivered prematurely. Screening for congenital hypothyroidism should be carried out after at least three days to avoid spurious results.

Infancy and childhood. Plasma concentrations of TSH are within the adult range but free T₃ is higher than in adults. Free T₄ concentrations tend to be at the lower end of the adult range. This hormone pattern suggests it may arise as a result of increased D1 activity in children. After puberty, no major changes in thyroid function occur, except in the pregnant woman.

Elderly. In old age, there is little change in thyroid function tests and age-related reference ranges are not required. There is a modest decrease in T₄ secretion but without an accompanying change in circulating T₄. A slight fall in T₃ may occur in those over 80 years of age. In patients on T₄ replacement, the dose of T₄ may have to be decreased with age. There is also increasing evidence that the normal TSH distribution curves appear to be shifted towards higher value ranges in individuals aged over 80. Age-specific analysis of TSH concentrations and anti-thyroid antibody titres measured in the National Health and Nutrition Examination Survey (NHANES) demonstrated that 12% of subjects aged 80 and older without any evidence of underlying autoimmune thyroiditis had TSH concentrations >4.5 mU/L. What is currently unclear is whether higher TSH concentrations confer a survival advantage or simply represent a degree of non-autoimmune age-related thyroid failure.

Pregnancy

In normal pregnancy there is a large increase in plasma TBG concentration that arises from both an oestrogen stimulated increase in synthesis and diminished clearance

of the protein. There is also an increase in the pool size of extrathyroidal T4 distribution and an increase in the deiodination of thyroid hormones in the developing placenta. The result of these changes is that, during pregnancy, there is a marked increase in the requirement for iodine and an approximately 50% increase in T4 production occurs if the supply of iodine is adequate.

In order to maintain homeostasis, an increase in total T4 and total T3 concentrations occurs, which reach a new steady state around mid-gestation. This requirement for increased thyroid hormone production requires an ideal iodine supply during pregnancy of approximately 250 µg/day. Similarly, women with hypothyroidism taking T4 replacement will need an increase in the dose of T4 of approximately 25–50 µg/day during pregnancy within the first six weeks of gestation, aiming for a TSH no higher than 2.5 mU/L. Some centres recommend that women double their pre-pregnancy levothyroxine dose on two days a week as soon as pregnancy is confirmed to reduce the risk of maternal hypothyroidism. Thyroxine replacement should be monitored carefully using both TSH and free T4, since even modest degrees of hypothyroxinaemia in early pregnancy have been associated with an impaired IQ in the infant.

In pregnancy, free T4 and free T3 concentrations may initially show a slight rise, thought to be due to the weak thyroid stimulating action of hCG, present in very high concentration in early pregnancy. This increase in free hormones can lead to suppression of TSH, such that low or sometimes undetectable concentrations may be found in up to 20% of patients during the first trimester. It is important that these changes in early pregnancy are not construed as suggesting that the patient has thyrotoxicosis. As pregnancy progresses, the concentrations of free thyroid hormones fall and TSH begins to rise, although rarely increasing above the reference range seen in non-pregnant subjects. Free thyroid hormone concentration may be lower than those found in non-pregnant women. It is important that trimester-related reference ranges are used for both TSH and free thyroid hormones. Patients can usually return to their pre-pregnancy levothyroxine replacement regimen immediately post-delivery; however, they should be monitored to ensure that this is still the optimal dose.

Hyperemesis gravidarum (severe vomiting in the first trimester of pregnancy) is associated with high plasma total and free thyroid hormone and low TSH concentrations, making it difficult to distinguish this condition from severe true thyrotoxicosis (thyroid crisis). It is believed that hCG is responsible for the thyroid stimulation; the condition usually resolves by the second trimester. Thyroid stimulating hormone receptor antibodies are negative in patients with hyperemesis but positive if the patient has Graves disease.

Thyrotrophin releasing hormone can cross the placenta from mother to fetus, but TSH does not. Thyroid hormones cross the placenta, but this transplacental flux is regulated in part by changes in expression of placental deiodinase as pregnancy progresses. A maternal supply of thyroid hormones (particularly T4) to the fetus is particularly important in the first trimester until a functional fetal thyroid has developed. However, since a fetus with

congenital thyroid deficiency develops relatively normally in utero, this suggests that some maternal-fetal transfer of thyroid hormone must be maintained throughout pregnancy. Iodide and antithyroid drugs also pass to the fetus. Indeed, antithyroid drugs are transferred more readily than thyroid hormone, and for this reason treatment of thyrotoxicosis in pregnancy with a 'block and replace' regimen (high dose antithyroid drugs 'rescued' by simultaneous thyroid hormone replacement) is contraindicated in pregnancy as it would result in a relatively hypothyroid fetus. The treatment of hyperthyroidism in pregnancy requires careful consideration. Radioactive iodine is contraindicated and the choice of antithyroid drug is currently the subject of debate. Propylthiouracil has been traditionally preferred to carbimazole/methimazole owing to fears of scalp and gastro-oesophageal defects in fetuses exposed to carbimazole/methimazole in the first trimester. However, recent concerns have arisen owing to the occurrence of propylthiouracil induced hepatitis, which has occasionally required liver transplantation and rarely has resulted in maternal death. Autoantibodies may cause additional problems in managing patients with hyperthyroidism in pregnancy. In patients with Graves disease, the transplacental passage of thyroid stimulating immunoglobulins may cause transient neonatal thyrotoxicosis, while the presence of anti-TPO antibodies is associated with an increased risk of maternal hypothyroidism and miscarriage.

Non-thyroidal illness

Patients attending or admitted to hospital suffering from any of a wide range of chronic or acute non-thyroidal illnesses (NTI), often have abnormalities in thyroid function tests. A low T3 may often be found even though the patients are clinically euthyroid; this has been termed the *sick euthyroid syndrome*.

Several mechanisms are involved, including:

- alterations in the hypothalamic–pituitary–thyroid axis leading to decreased hypothalamic stores of TRH and suppression of TSH release due to increased concentrations of dopamine, cytokines, cortisol and somatostatin (see Fig. 19.6). While the degree of TSH suppression in NTI is often not as great as that found in hyperthyroidism, there is considerable overlap in TSH concentrations found in the two conditions
- changes in the affinity characteristics and in the plasma concentrations of the thyroid hormone-binding proteins. These changes give rise to alterations in the plasma concentrations of both the free and total thyroid hormones
- impaired uptake of thyroid hormones by the tissues
- decreased production of T3 in the peripheral tissues
- changes in the T3 occupancy and function of the T3 receptors.

The contribution of each of the above mechanisms may vary with the severity and stage of the illness, and thus the pattern of thyroid function tests may be extremely variable and may mimic the profile seen in either primary or secondary thyroid disease. Interpretation of

thyroid function tests is complicated further by the effects of drugs and methodological problems associated with free hormone measurements.

Thyroid hormone metabolism is markedly affected by fasting and illness, with the magnitude of these changes tending to be proportional to the severity of the illness (Fig. 19.7). Extrathyroidal conversion of T₄–T₃ is reduced, leading to a marked decline in plasma total and free T₃, which may fall to undetectable concentrations. Reverse T₃ concentration increases, primarily owing to impaired catabolism rather than increased synthesis. These changes are often considered to be an adaptive response for energy conservation, as rT₃ is not metabolically active. Concentrations of free fatty acids and other substances that can compete with thyroid hormones for binding to plasma proteins may rise. This can produce a transient increase in free T₄. Drugs that compete for T₄ binding sites will have a similar effect. Uptake of thyroid hormone into cells may be impaired, either directly by endogenous inhibitors or indirectly as a result of impairment of the active transport systems of the cells. Finally, administration of corticosteroids or dopamine may suppress TSH release.

Methodological shortcomings render the results of free hormone analysis liable to serious error. Equilibrium dialysis methods usually show a normal or slightly raised free T₄ in patients with illness of mild to moderate severity, but most commercial assays tend to show normal or low free T₄ results in sick patients. Normalization or transient rebound changes of thyroid parameters may be seen during recovery from NTI or refeeding after starvation. In particular, TSH may rise transiently into the hypothyroid range. It should be noted that, although TSH is in general the most reliable test of thyroid function, in hospitalized patients a TSH of <0.1 mU/L is twice as likely to be due to NTI as to hyperthyroidism, and an increased TSH is as likely to be due to recovery from illness as to hypothyroidism. Because of the poor predictive value of thyroid function tests in hospitalized patients, these tests should only be requested if the reason for hospital admission is considered, on clinical grounds, to be related to a thyroid problem. Also, abnormal thyroid function tests during illness may need to be rechecked after the acute illness has resolved.

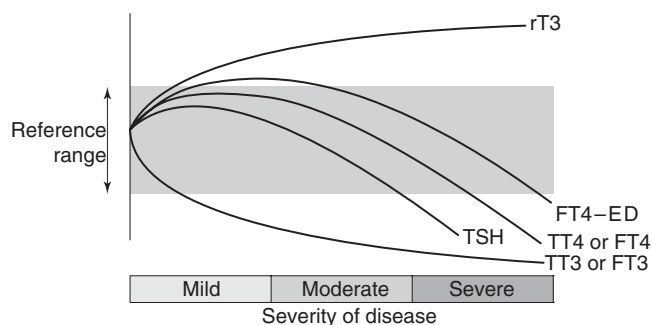


FIGURE 19.7 ■ The effects of illness on the concentration of thyroid hormones and thyrotrophin. The profile for free T₄ obtained using equilibrium dialysis is shown (FT4-ED). The concentration of free T₄ found using routine assays is method dependent, with some assays showing the profile of FT4-ED, while others follow the profile of total T₄ (TT4). Free T₃ (FT3) shows a similar profile to total T₃ (TT3).

While the changes in thyroid hormone metabolism associated with illness might theoretically lead to tissue hypothyroidism, most trials of T₄ and T₃ administration in severe non-thyroidal illness have not shown improved outcome.

Drugs

Drugs may interfere with TSH secretion or the production, secretion, transport and metabolism of thyroid hormones (Table 19.1). Some drugs modify thyroid status while others produce abnormal thyroid function test results in clinically euthyroid subjects. Certain agents (in particular iron preparations) impair the absorption of T₄ from the gut, and patients on thyroxine therapy should be advised to take their T₄ at least 4 h apart from these medications. Patients should also be advised to avoid caffeine within 45 min of taking levothyroxine as caffeine can also impair absorption. Patients taking levothyroxine are likely to require an increase in replacement dose if drugs such as phenytoin or carbamazepine are also prescribed as they increase hepatic metabolism of T₄. Phenytoin, carbamazepine, furosemide and salicylate compete with thyroid hormone binding to plasma binding proteins and may increase plasma free T₄ concentrations. The influence of

TABLE 19.1 Drugs affecting thyroid function

Mechanism	Example of drug
Decrease in TSH secretion	Dopamine, glucocorticoids, octreotide, cytokines
Decrease in thyroid hormone secretion	Lithium, amiodarone, iodide
Increase in thyroid hormone secretion	Lithium, amiodarone (rare), iodide
Decrease in thyroidal synthesis ^a	Carbimazole, propylthiouracil, lithium
Increase in TBG	Oestrogens, tamoxifen, heroin, methadone, clofibrate, raloxifene
Decrease in TBG	Androgens, glucocorticoids, anabolic steroids
Displacement of thyroid hormones from plasma proteins	Furosemide, fenclufenac, salicylates, mefenamic acid, carbamazepine
Increased hepatic metabolism	Phenytoin, carbamazepine, rifampicin, barbiturates
Impaired T ₄ and T ₃ conversion	β-antagonists, amiodarone ^a , radiocontrast dyes, e.g. iopanoic acid
Impaired absorption of thyroxine ^b	Cholestyramine, aluminium hydroxide, ferrous sulphate, calcium salts, sucralfate, soya protein
Altered immunity ^c	Interleukin-1, interferons, tumour necrosis factor-α, interleukin-2

^aCause decrease in thyroid hormone synthesis or secretion and alter thyroid status.

^bInterfere with absorption from GI tract. Patients on T₄ therapy should be advised to take their T₄ at least 4 h apart from these medications.

^cThese cytokines can cause transient hypothyroidism or thyrotoxicosis, the mechanism of which is unclear.

Other drugs listed produce abnormal thyroid function tests but patients remain euthyroid. Amiodarone is the exception (see text).

drugs on modifying free thyroid hormone concentrations may be method specific. For example, depending on the method, free T4 may be measured as being normal, high or low in patients given heparin or taking phenytoin or carbamazepine. Amiodarone, lithium and interferon can induce thyroid dysfunction and are covered in detail later in this chapter.

THE EVALUATION OF THYROID FUNCTION

This section describes the methods that have been developed in order to assess the functional state of the thyroid gland. In most cases, the presenting clinical features will lead to thyroid function tests being performed for confirmation of the provisional diagnosis.

Clinical evaluation of thyroid status

Disordered thyroid function may be detected clinically by demonstrating both changes in the thyroid gland itself and the systemic effects of under- or overproduction of

thyroid hormones. In some instances, the clinical presentation may be virtually diagnostic of a specific condition and investigations are required only for confirmation. In others, the clinical features may be less specific and more thorough investigation is required. The signs and symptoms of thyroid dysfunction are listed in Tables 19.3 and 19.6. The normal thyroid can often be palpated and may also be apparent on inspection of the neck, particularly in young, thin women. Abnormal thyroid enlargements may be diffuse or asymmetrical. As a result of connective tissue attachments to the trachea, thyroid swellings move upwards on swallowing and a similar phenomenon is found with tongue protrusion in the case of thyroglossal cysts. The normal gland has a 'rubbery' feel; in Graves disease and diffuse colloid goitre, the consistency is softer. By contrast, the thyroid in Hashimoto disease is often rather firm, often with a palpable pyramidal lobe. Thyroid carcinoma (anaplastic) and Riedel thyroiditis result in a rock-hard gland that may, in the case of carcinoma, be irregular in outline. Increased blood flow through a hyperactive gland is often associated with a bruit. This may be maximal over the superior thyroidal

TABLE 19.2 Interpretation of thyroid function tests

	TSH low	TSH normal	TSH high
Free T4 low	Severe non-thyroidal illness: repeat test when non-thyroidal illness resolved	Non-thyroidal illness or drugs: repeat tests when non-thyroidal illness resolved	Hypothyroidism: no confirmation needed but look for cause—autoimmune, drugs, congenital etc.
Free T4 normal	Recent treatment for hyperthyroidism Hypopituitarism: other tests of pituitary function abnormal Subclinical hyperthyroidism: free T3 normal; check autoantibodies (antiperoxidase) Over-replacement with T4 Non-thyroidal illness, drugs: free T3 normal or low; only common in hospitalized patients T3-hyperthyroidism: free T3 high; check autoantibodies Treated thyrotoxicosis: free T3 normal or high	Recent treatment for hyperthyroidism Hypopituitarism: other tests of pituitary function abnormal Euthyroid: no confirmation needed. Also applies to adequate T4 replacement	Inadequate T4 replacement Subclinical hypothyroidism: check autoantibodies Inadequate T4 therapy Recovery from non-thyroidal illness; repeat tests after illness fully resolved
Free T4 high	Hyperthyroidism: free T3 usually high; check autoantibodies Over-replacement with T4 Hyperthyroidism with coexistent non-thyroidal illness: total T4 raised, free T3 may be normal; check autoantibodies Hyperemesis gravidarum or very early pregnancy: free T3 normal, sometimes raised; check TSH receptor antibodies to exclude Graves disease Non-thyroidal illness: total T4 usually normal; free T3 usually normal or low	T4 replacement: either erratic compliance with T4 replacement therapy or consistent with adequate replacement in some patients Non-thyroidal illness, drugs Autoantibodies to T4 (and T3): free T3 may be high; test for autoantibodies to thyroid hormones; try a different method for free T4 Genetic albumin variants: try a different method for free T4. Resistance to thyroid hormones: free T3 high; SHBG normal; normal or exaggerated response of TSH to TRH; family studies TSH-secreting tumour: free T3 high; SHBG high; no TSH response to TRH or T3; other tests of pituitary function abnormal; α -subunit increased	Erratic compliance with T4 therapy Resistance to thyroid hormones: free T3 high; normal or exaggerated response of TSH to TRH; family studies TSH-secreting tumour: free T3 high; no response of TSH to TRH or T3

Bold type shows most common diagnoses in each category.

artery rather than the gland itself and, in some cases, may also produce a palpable thrill.

In vitro tests of thyroid activity and pituitary–thyroid status

The tests used to investigate thyroid dysfunction can be grouped into:

- tests that establish whether there is thyroid dysfunction, e.g. serum TSH and thyroid hormone (T4 and T3) measurements
- tests to elucidate the cause of thyroid dysfunction, e.g. thyroid autoantibody and serum thyroglobulin measurements, thyroid enzyme activities, biopsy of the thyroid, ultrasound and isotopic thyroid scanning
- thyroglobulin measurements, which are used to monitor treatment and detect recurrence in patients with follicular or papillary carcinoma.

Measurement of TSH and thyroid hormones should be performed to determine the patient's thyroid status before requesting the additional tests that seek to determine the cause of any thyroid dysfunction.

Measurement of thyroid stimulating hormone

The measurement of TSH in a basal serum sample by a sensitive immunometric assay provides the single most sensitive, specific and reliable test of thyroid status in both overt and subclinical primary thyroid disorders. In primary hypothyroidism, TSH is increased, while in primary hyperthyroidism TSH is usually <0.02 mU/L. However, TSH alone is not a reliable test for detecting thyroid dysfunction arising from hypothalamic–pituitary dysfunction and in other specific instances as outlined below.

Thyroid stimulating hormone is now measured by immunometric assays (IMAs) that utilize non-competitive labelled antibody methods with non-isotopic labels. These IMAs use highly specific monoclonal antibodies raised to epitopes on α - and β -subunits of the TSH molecule. The IMA technique uses a sandwich-assay configuration using two antibodies. The 'functional sensitivity' of a TSH assay defines the minimum concentration of TSH that can be quantified in routine use. To obtain the functional sensitivity of a routine assay, a precision profile is constructed using data obtained over many assay runs with multiple batches of reagent and different operators. The concentration of TSH that has a coefficient of variation of 20% taken from the precision profile defines 'functional sensitivity'. It is essential that TSH assays are sufficiently sensitive for clinical diagnostic purposes and they must have a functional sensitivity of <0.02 mU/L.

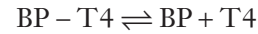
Free T4 and free T3 measurements

Theoretical considerations

The plasma concentrations of free thyroid hormones are extremely low and their quantification in the presence of large concentrations of bound hormones has proved challenging. Poor assay design of many free

hormone methods has produced a mass of conflicting literature regarding the typical ranges of free thyroid hormone concentrations in patients with NTI and taking a variety of drugs.

The free hormone concentration results from the equilibrium between the bound (BP) and free hormone.



The proportion of T4 that binds to the binding protein is dictated through the law of mass action by the affinity of the binding protein (K_{eq}) multiplied by its concentration. This is known as the 'relative binding capacity'. Thus, if the affinity or concentration of the binding protein decreases, the proportion of T4 bound will also diminish and the free hormone concentration will increase. In humans, the free T4 fraction is mostly affected by changes in the affinity and concentration of TBG; changes in the affinity and concentration of albumin or transthyretin have little effect on free T4.

Free hormone measurements involve sampling the free hormone fraction. This can be done using physical separation of the free from bound hormone using a semipermeable membrane (equilibrium dialysis or ultrafiltration) or, alternatively, by adding an antibody that 'captures' a proportion of the free hormone pool. The removal of free hormone from the original equilibrium will result in further hormone dissociating from the binding protein to create a new equilibrium. A crucial requirement of any valid method for free hormone measurement is that during the assay process there should be minimal disruption of the original equilibrium, and this can be achieved by minimizing the dilution of the specimen and sampling only a small proportion of the hormone pool. The pH and the ionic composition of any buffer also need to be controlled. The inadequacies of many commercial assays have led to the suggestion that the term *free hormone estimate* is used for all free hormone assays used in clinical practice.

Methods for measuring free thyroid hormones

Equilibrium dialysis and ultrafiltration methods

These are widely regarded as reference methods, although neither is completely satisfactory. The long dialysis period (~12–18 h) used for equilibrium dialysis can, in certain samples (e.g. patients given heparin), lead to the production of non-esterified fatty acids (NEFAs) from triglyceride, which may displace thyroxine from its binding proteins and thus increase the free T4 concentration. A similar increase in free T4 will occur if such patient samples are stored unfrozen prior to analysis.

Validity of commercial methods for free hormone analysis

Routine assays are given a variety of names depending on the methodology used, but all usually involve: (1) adding an antibody to diluted serum to 'capture' a proportion of the free T4 pool and (2) estimating the

unoccupied T4 binding sites on the antibody by adding labelled T4. These commercial assays have additional problems over those of equilibrium dialysis. First, the antibody used will tend to disrupt the original equilibrium, an effect that is dependent on the overall thyroxine binding capacity of the sample and the affinity and concentration of the antibody. Second, the binding capacity of the patient sample may be quite different to that used in the assay calibrators. The overall effect of this is that samples with low binding capacity (NTI, low TBG etc.) tend to have a negative bias introduced, while samples with high binding capacity (high TBG, e.g. pregnancy) may have a positive bias introduced. Many commercial assays include albumin, added in an attempt to reduce non-specific binding and nullify the effects of NEFA on free T4 concentrations. Unfortunately, albumin binds T4 and, if present in sufficient concentrations, will cause significant disruption of the equilibrium. Again, samples with low serum binding capacity will be more prone to this effect than samples with normal or high binding capacities. These various effects explain why free T4 assays may appear to produce low, normal or raised free T4 concentrations in the same sample taken from patients with NTI.

In a true free hormone assay the calculated and measured free hormone concentrations in a set of serially diluted serum samples should be closely approximate. Assays that show a significant reduction in free T4 at these dilutions are likely to produce heavily negatively biased results in patients with low serum binding capacity. Many current commercial assays for free T4 fail these validation tests. The development of free T3 assays has been even more problematic than for free T4.

Nomenclature of free thyroid hormone assays

The names given to free hormone assays include one-step analogue, two-step back-titration and labelled antibody assays. The principles for each of these assays are widely reviewed in the literature but the name given to an assay does not guarantee its validity or performance.

The free T4 and free T3 index methods, that estimate the free hormone concentration in the presence of protein-bound hormone, have now largely been replaced by the direct measurement of free T4 and free T3.

Total T4 and total T3

Plasma concentrations of total T4 and total T3 are measured by immunoassay using diluents that incorporate a chemical agent that releases the thyroid hormones from their binding proteins. In clinical practice, most problems associated with total hormone measurements arise in patients with modified hormone binding capacities that, in turn, may increase or decrease the total hormone concentration. Pregnancy and oestrogen medication give rise to increased concentrations of TBG and thus increases in total T3 and total T4. Some causes of abnormal TBG concentration (and thus, abnormal total thyroid hormones) are given in [Box 19.2](#). Genetic variants of TBG, transthyretin and albumin have been described that have

BOX 19.2 Causes of abnormal plasma thyroid binding globulin (TBG) concentrations

High TBG

- Genetic (hereditary TBG excess X-linked dominant)
- Physiological (pregnancy, neonatal period)
- Hydatidiform mole
- Oestrogens, including oestrogen-containing oral contraceptives, tamoxifen,
- Other drugs, e.g. phenothiazines, opiates, clofibrate
- Hepatitis
- Acute intermittent porphyria

Low TBG

- Genetic
- Androgens
- Protein-losing states, e.g. nephrotic syndrome
- Malnutrition
- Malabsorption
- Severe illness
- Acromegaly
- Cushing disease
- Corticosteroid therapy (high dose)
- Liver disease (chronic)

altered hormone-binding characteristics, which in turn modify the total concentration of thyroid hormones in the circulation. Non-thyroidal illness results in the production of modified TBG with diminished affinity for thyroid hormone. Many drugs may also lower the hormone binding capacity of plasma. Some patients produce autoantibodies to the thyroid hormones that increase the hormone binding capacity of plasma and, in addition, produce marked assay interference.

The use of total hormone measurements is now less popular following the introduction of more reliable methods for free hormone measurements. However, total hormones are still considered to have a role in clarifying thyroid status in some patients with NTI where the results of free thyroid hormone and TSH measurements do not concur.

Selective use of thyroid function tests

Many laboratories measure random TSH as the initial test of thyroid function. This approach is not infallible, but is least likely to produce an abnormal result in NTI and in patients on various forms of drug treatment. It can detect both overt and subclinical disease. A normal TSH concentration effectively excludes *primary* thyroid dysfunction. If an abnormal result is obtained, thyroid hormone measurements are then made to confirm that thyroid dysfunction is present, and to determine the severity of the disease.

Initial measurement of both TSH and free T4 together can provide a more satisfactory method of assessing thyroid status, since in some situations a single TSH result may be misleading. If this strategy is followed, a significant number of cases will arise in which one test is abnormal while the complementary test is normal. It is thus essential to understand and appreciate the factors that can affect the results of thyroid function tests.

Interpreting results of thyroid function tests

Table 19.2 provides a guide to the interpretation of thyroid function tests.

Situations in which TSH usually provides the correct estimate of thyroid status

Overt primary hyperthyroidism. Thyroid stimulating hormone is nearly always $<0.1\text{mU/L}$ and often $<0.02\text{mU/L}$, owing to feedback inhibition on the pituitary. Free and total T4 and T3 concentrations are nearly always increased, but in a very small percentage of hyperthyroid patients, total T4 and free T4 are both normal, whereas both total T3 and free T3 are increased: this condition is known as T3 hyperthyroidism or T3 thyrotoxicosis.

Overt primary hypothyroidism. Thyroid stimulating hormone is invariably increased, often to $>20\text{mU/L}$, as feedback inhibition of the pituitary is diminished, and free T4 and total T4 are low. Free T3 and total T3 measurements are of no value here, since normal concentrations are often observed because of increased peripheral deiodination of T4–T3.

Subclinical thyroid disease. Thyroid disease presents as a spectrum of clinical and biochemical features of varying severity. The clinical diagnosis of mild thyroid disorders is often difficult, and the only biochemical abnormality may be an abnormal serum TSH concentration. For example, many clinically euthyroid patients with multinodular goitre or with exophthalmic Graves disease have a TSH $<0.1\text{mU/L}$, but normal free and total thyroid hormone concentrations. The combination of a low TSH together with normal thyroid hormone concentrations in a patient is known as ‘subclinical hyperthyroidism’, but this description is unsatisfactory, since it rests solely on the results of chemical investigations. Similarly, a diagnosis of ‘subclinical hypothyroidism’ is given to those patients with an increased TSH but normal thyroid hormone concentrations.

Before the diagnosis of subclinical thyroid disease can be made, causes of an abnormal plasma TSH other than thyroid disease must be excluded. These include pregnancy, NTI, drug treatment and assay interference.

Common situations in which TSH results may be misleading

Assay interference from endogenous heterophilic antibodies

Some individuals have antibodies in their plasma that react with a range of animal immunoglobulins (heterophilic antibodies). These antibodies interfere with a wide range of immunoassays. For example, a normal or even elevated TSH result may be found in some thyrotoxic patients owing to this type of assay interference.

Pregnancy

Thyroid stimulating hormone is a reliable indicator of thyroid status in the second and third trimester of

pregnancy, but in the first trimester, a low TSH may be found in up to 20% of women. Thyroid function testing in pregnancy is covered earlier in this chapter.

Secondary thyroid disorders

Thyroid stimulating hormone concentration is normal in about half of patients with central (pituitary) hypothyroidism, and in a few, TSH may be slightly elevated. Circulating TSH has been shown to have reduced bioactivity in central hypothyroidism and thus free T4 and total T4 concentrations are usually low. When pituitary disease is known to be present, TSH concentrations should not be used to assess the thyroid status.

Very rarely, hyperthyroidism is due to a TSH-secreting tumour. Persistent hyperthyroid symptoms associated with elevated free T4 and free T3 but raised or normal TSH are consistent with this diagnosis, once the common problems of assay interference or NTI have been eliminated. Serum α -subunit concentrations are often increased in these patients.

Abnormal pituitary T3 receptor function, due generally to mutations in the thyroid hormone β receptor, can result in disordered thyroid function not reflected by changes in TSH. In this disorder, the thyrotroph is resistant to the normal negative feedback regulation, so that plasma thyroid hormone concentrations are raised in the presence of a normal or raised TSH. Depending on the balance of effects between receptor function in the pituitary and the rest of the body, the patient may be clinically euthyroid, hyperthyroid or even mildly hypothyroid. This condition is termed ‘thyroid hormone resistance’.

Reference ranges and significant changes

By convention, a reference range usually only comprises 95% of values from a reference population; therefore 2.5% of the population will have values above the reference range and 2.5% below. Thyroid hormone concentrations show a normal distribution in reference populations. For TSH, the reference population shows a log normal distribution even when subjects who are thyroid peroxidase (TPO) antibody positive are excluded. Thyroid stimulating hormone reference ranges should be established using specimens collected between 08.00 h and 18.00 h and using 95% confidence limits from log transformed data. The reference population should have no personal or family history of thyroid dysfunction, be on no medication known to alter TSH secretion and have no thyroid antibodies detectable by a sensitive assay. Method-related reference ranges need to be used for all thyroid function tests. Typical adult reference ranges are: TSH 0.3–4.0 mU/L, total T4 60–160 nmol/L, total T3 1.2–2.6 nmol/L, free T4 9–25 pmol/L and free T3 3–8 pmol/L.

In neonates, infants and children, age-related reference ranges are recommended. Trimester-related reference ranges should be available for use in pregnancy.

Knowledge of the analytical variability and estimates of the intra-individual and inter-individual variability allows a calculation of differences in the results of thyroid tests that can be considered as clinically significant

when monitoring serially a patient's response to therapy. These changes may be method dependent but are in the order of:

TSH	0.8 mU/L
Total T4	15 nmol/L
Free T4	6 pmol/L
Total T3	0.6 nmol/L
Free T3	1.5 pmol/L

Miscellaneous tests

Thyrotrophin releasing hormone test

In this dynamic test of pituitary-thyroid function, the TSH response to an intravenous dose of thyrotrophin releasing hormone (TRH) is measured. Formerly, it was used as a test of hypothyroidism when TSH estimates were insensitive (exaggerated TSH response) and as a test of pituitary function. Nowadays it is only occasionally used in the diagnosis of thyroid hormone resistance or TSH-secreting adenoma.

Thyroglobulin

Thyroglobulin (Tg) is normally present in the circulation in very small amounts and, although concentrations may be raised in many thyroid disorders, its measurement (by immunoassay) is essentially reserved for follow-up of patients with thyroid cancer (post total thyroidectomy and radioiodine ablation), in whom elevation of previously suppressed concentrations may indicate tumour recurrence. The measurement of Tg is challenging, as up to 30% of patients with thyroid cancer have antithyroglobulin antibodies in their serum that may interfere with the assay. Rarely, demonstration of suppressed Tg concentration in the presence of raised T4 and T3 is used to diagnose factitious hyperthyroidism due to illicit abuse of thyroid hormone.

The measurement of serum thyroglobulin is also useful in identifying the cause of congenital hypothyroidism as all patients with organification disorders, and those with severe thyroid-stimulating hormone receptor defects, will have raised concentrations when their circulating TSH is increased, except for those with thyroglobulin synthesis defects where the thyroglobulin concentration usually remains low. Thyroglobulin measurements can also provide useful information about the presence of thyroid tissue, for example, if serum thyroglobulin is measurable where there is apparent athyreosis on imaging, this implies that radiologically undetectable hypoplastic thyroid tissue is present.

α -Subunit

The pituitary glycoprotein hormones TSH, follicle stimulating hormone (FSH) and luteinizing hormone (LH) are composed of two subunits, the α -subunit being common to all three hormones and the β -subunit specific to each hormone. The plasma concentration of the α -subunit can be measured by immunoassay, without cross-reaction from complete hormone molecules, and it

may be raised in patients with pituitary tumours. Raised concentrations are also found in some postmenopausal women.

Autoantibodies to thyroidal antigens

Antibodies to thyroid peroxidase (TPOAb)

Originally known as thyroid microsomal antibodies, these antibodies are found in almost all (95%) patients with autoimmune hypothyroidism secondary to Hashimoto thyroiditis early in the course of the disease, and in some patients with other autoimmune thyroid diseases. The antibodies are polyclonal and their target antigen has been identified as the thyroid peroxidase enzyme, which plays critical roles in thyroid hormone synthesis. These antibodies are frequently present in individuals with subclinical hypothyroidism as well as in 2–5% of biochemically euthyroid women (rising to over 10% over the age of 70). The prevalence of these antibodies is lower in men, but also increases with age, being present in 2.5% of the male population aged over 75. They may play a pathogenetic role in some patients with destructive autoimmune thyroid disease, as they have been shown both directly to fix complement on the thyrocyte and to direct cell-mediated cytotoxic reactions against thyrocytes, but in the majority of cases the damage to the thyroid is believed to be T cell rather than antibody mediated.

Knowledge of the presence of this group of antibodies can be of value in several circumstances. The presence of a high titre of TPOAb in the presence of goitrous hypothyroidism is characteristic of Hashimoto disease; they may also be seen in association with the early, transient hyperthyroidism sometimes associated with this condition. In pregnancy, the presence of antithyroid peroxidase antibodies is a significant risk factor for miscarriage, maternal hypothyroidism and the future development of postpartum thyroiditis with disturbed thyroid function. In more general terms, the finding of TPOAb in high titre is indicative of thyroid autoimmunity and may hence be associated with present or future occurrence of other organ-specific autoimmune disease. Patients with ophthalmic Graves disease (eye signs typical of Graves ophthalmopathy but without a present or past history of hyperthyroidism) are often found to have these antibodies, confirming the association with autoimmune thyroid disease despite normal thyroid function. In patients with subclinical hypothyroidism, the presence of TPOAb is an indicator to consider commencing T4 replacement, as such patients are at significant risk of developing overt hypothyroidism in the future (50% will progress to biochemical hypothyroidism over the next 20 years).

Antibodies to thyroglobulin (TgAb)

Antibodies to thyroglobulin are also found in patients with thyroid autoimmunity, but at lower frequency than the TPOAb. These antibodies do not fix complement and are not known to play a direct pathogenetic role in the aetiology of autoimmune thyroid disease in man. The only reason to measure these antibodies is in patients

being treated for follicular or papillary cancer, to indicate possible assay interference in Tg assays. Antibodies to thyroglobulin can interfere with assays for Tg, giving falsely elevated Tg concentrations in radioimmunoassays and falsely decreased concentrations in IMAs. In patients being monitored for differentiated thyroid cancer, a rising titre of TgAb is associated with a poor prognosis.

Antibodies to the thyroid stimulating hormone receptor

The hyperthyroidism of Graves disease is mediated via the effect of IgG class autoantibodies that bind to the TSH receptor and produce an increase in intracellular cyclic AMP concentration. Serum concentrations of such antibodies can be correlated with thyroid gland hyperfunction. Two types of assay system are in use: the first measures the capacity of the autoantibody to inhibit the binding of labelled TSH to its receptor on human thyroid membrane preparations (TSH-binding inhibiting immunoglobulin or TBII), and the second measures the increase in cyclic AMP concentrations in thyroid preparations following stimulation by the patient's IgG (thyroid stimulating antibody or TsAb). During periods of thyroid overactivity, these assays will be positive in 90–95% of patients.

Knowledge of TSH receptor antibody status in patients with Graves disease during pregnancy is of prognostic value: the infants of patients with a high antibody titre during the third trimester have an increased risk of neonatal hyperthyroidism secondary to transplacental passage of maternal autoantibody. Such neonatal hyperthyroidism is transient, settling as maternal antibody disappears from the child's circulation.

Other TSH receptor antibodies ('blocking' antibody), although infrequently demonstrated, can inhibit gland function and lead to hypothyroidism; the predominant effect (stimulation vs blocking) may vary in a single patient with time. However, stimulation is far more commonly found.

Antibodies and ophthalmopathy of Graves disease

The aetiology of the ophthalmopathy of Graves disease, which is present in about a third of patients, is still poorly understood, but current evidence suggests an autoimmune pathogenesis and an association with smoking. Orbital muscle, connective tissue and adipose tissue become infiltrated with lymphocytes and macrophages. The extracellular compartment of the extraocular muscles and orbital fibro-adipose tissue becomes oedematous owing to water deposition caused by production of glycosaminoglycans by orbital fibroblasts and pre-adipocytes. Evidence of autoimmune responses to thyroid antigens is often present in patients with Graves ophthalmopathy. The most promising candidate autoantigen is the TSH receptor that is expressed in orbital connective tissue and in particular orbital fat (pre-adipocytes). It seems most likely that TSH receptor binding antibodies activate orbital fat cells by signaling through this receptor in a way that is subtly different from activation of thyroid

cells, explaining why Graves ophthalmopathy is often, but not always, associated with stimulation of the thyroid gland. A role for activation of the insulin-like growth factor (IGF-1) receptor in orbital fat has also been proposed.

Imaging the thyroid

Standard radiography gives only limited information in the case of thyroid enlargement and is not used routinely. X-ray of the chest may show a superior mediastinal mass in patients who have retrosternal extension of a goitre, and the presence of tracheal deviation with or without narrowing of the lumen may also be noted.

Thyroid ultrasound can demonstrate the type of thyroid enlargement (diffuse or localized) and nature (solid or cystic, single or multiple) of thyroid nodules. It is, however, not possible definitively to exclude malignancy using ultrasound alone and biopsy of the nodule may be required. There is increasing use of ultrasonography to improve the quality of the content of fine needle aspirations of thyroid nodules. Colour-flow Doppler sonography is also increasingly being used to distinguish between type 1 amiodarone-induced thyrotoxicosis (increased uptake) and type 2 amiodarone-induced thyrotoxicosis (decreased/absent uptake); these conditions are covered later in this chapter.

The ability of the thyroid to concentrate radioisotopes of iodine and other substances has been utilized for many years to identify areas within the gland that are or are not functionally active. These imaging systems are of particular value in the evaluation of thyroid nodules, where the presence of nonfunctioning solid tissue is associated with a much higher risk of malignancy. Although subject to important limitations, the level of overall gland activity can also be assessed.

Thyroid scintiscanning

Using intravenous administration of radioisotopes and a scintillation camera, qualitative images of functioning thyroid tissue can be obtained and the proportion of the total dose taken up by the gland can be quantified. Of the available isotopes, ^{99m}Tc , given intravenously as pertechnetate, is concentrated within the gland but not organified into thyroid hormones and therefore diffuses out of the gland with time. This fact, together with its short half-life, means that large doses can be administered without delivering a high radiation dose to the thyroid. The gland is imaged 20 min after injection.

Disadvantages of use of this isotope are that it cannot reliably be used to identify retrosternal glands (as confusion may arise with intravascular radioactivity) and the lack of information regarding iodine organification. The use of radioisotopes of iodine circumvents this latter problem, ^{123}I perhaps being the best isotope to use when imaging thyroid tissue in its normal site as, by comparison with other isotopes of iodine, it delivers a much lower total radiation dose to the gland. Iodine-131, with higher energy and a longer half-life, is useful where deep thyroid tissue is sought or where there is a need to detect, and treat by irradiation, functionally active metastatic thyroid carcinoma.

By quantifying the proportion of the total administered dose of isotope concentrated within the thyroid gland during a given time period, it is possible to estimate the activity of the gland, uptake increasing with increased gland activity and vice versa. However, gland uptake of isotopes of iodine may be critically influenced by many factors other than gland activity (such as dietary iodine content, coincidental administration of iodine-containing drugs such as amiodarone and previous administration of iodine-containing contrast media for radiological investigations) and the results of such uptake studies should be interpreted with caution. Biochemical tests of thyroid function have therefore superseded uptake studies in the measurement of gland activity. However, uptake studies are useful in identifying those patients with hyperthyroidism and a negligible uptake (painless thyroiditis, previous treatment with amiodarone) for whom ^{131}I therapy would be inappropriate.

Perchlorate discharge test

Rarely, hypothyroidism and/or goitre result from defects in enzymes responsible for the organification of iodine into thyroid hormone. In such cases, iodine will be trapped within the thyrocyte but is not organified. Perchlorate, given orally as potassium perchlorate, will cause inhibition of the active transport of iodine across the thyrocyte membrane and hence lead to a loss of non-organified iodine from the gland.

HYPERTHYROIDISM

Clinical features

The clinical symptoms and signs which may result from the hyperthyroid state are summarized in Table 19.3. The clinical picture may be modified by the presence of coexisting systemic diseases and may depend on the age of the patient. Some of the thyroid diseases themselves also produce characteristic physical signs, for example the orbital and cutaneous manifestations of Graves disease. The changes seen in the major organ systems are described below.

Cardiovascular system

Many of the manifestations of hyperthyroidism relate to the increased demands placed upon the cardiovascular system. High circulating concentrations of thyroid hormones have a direct stimulatory effect on cardiac muscle. Heart rate and stroke volume are both increased at rest (including during sleep) and peripheral vascular resistance is reduced, leading to a marked rise in cardiac output in patients who have no pre-existing cardiac disease. Overt cardiac failure may result from this high output state but can also be triggered by the occurrence of arrhythmias. Atrial flutter/fibrillation is seen in over 10% of patients with hyperthyroidism and may be the presenting feature in some cases. Patients who suffer from ischaemic chest pain typically find that angina worsens as a result of the increased metabolic demands placed on the myocardium.

TABLE 19.3 The symptoms and signs of hyperthyroidism

	Common	Uncommon	
Symptoms	Increased irritability	Increase in appetite	
	Increased sweating	Leg swelling	
	Heat intolerance	Menstrual irregularity	
	Palpitation	True diarrhoea	
	Lethargy	Loss of appetite	
	Loss of weight	Weight gain	
	Breathlessness		
	Increased bowel movement		
	Signs	Tachycardia	Palmar erythema
		Goitre	Finger clubbing
Warm, moist peripheries		Splenomegaly	
Tremor		Pretibial myxoedema	
Arrhythmias		Gynaecomastia	
Eye signs			
Proximal myopathy			
Thyroid bruit			
Cardiac failure			
Onycholysis			
Muscular weakness			

Thyroid crisis

Florid hyperthyroidism may result in a life-threatening illness (thyroid 'crisis' or 'storm'). In these – fortunately uncommon – cases, cardiovascular symptoms and signs predominate. Circulatory collapse is characteristic and results from arrhythmia and high output cardiac failure. High output cardiac failure usually occurs in younger individuals with longstanding thyrotoxicosis; despite doubling of the cardiac output, the increase in preload and blood volume in thyrotoxicosis causes a rise in ventricular filling pressure leading to pulmonary and peripheral congestion. Pharmacological attempts to correct abnormal cardiac rhythms are usually unsuccessful in the face of continuing hyperthyroidism. Management aims to control the hyperthyroidism rapidly (using large doses of antithyroid drugs, including iodine) while treating the patient symptomatically. β -Blocking drugs such as propranolol are used, and measures to cool the patient are often needed. Use of steroids under these circumstances may be of additional value.

Gastrointestinal system

Weight loss is a classic feature of hyperthyroidism, often occurring despite an increased appetite. Increases in gastrointestinal motility lead to increased frequency of bowel movement, though true diarrhoea does not usually result. Nausea and vomiting are also uncommon, but may precede the onset of a thyrotoxic crisis. In severe hyperthyroid states, liver function can be markedly deranged, with hypoalbuminaemia and elevation of plasma aminotransferase and alkaline phosphatase activities. Paradoxically, alkaline phosphatase concentrations (of mixed bone and liver origin) can rise further initially after treatment for thyrotoxicosis, settling after 2–3 months.

Central and peripheral nervous system

Generalized hyperkinesia is very often seen in hyperthyroid individuals, frequently being accompanied by emotional lability. A fine tremor of the outstretched fingers is also characteristic. Patients complain of inability to sleep, despite often profound feelings of tiredness.

Locomotor system

Muscular weakness particularly affects proximal muscles. The aetiology of the weakness may be related to impaired phosphorylation of creatine. Muscle biopsy shows loss of type IIb muscle fibres, with replacement by fat and occasional lymphocytic infiltration. These changes reverse when euthyroidism is restored, but muscle power can take several weeks or months to return fully to normal. Periodic paralysis, associated with hypokalaemia during attacks of weakness, is also seen in association with hyperthyroidism, particularly in people of oriental descent. Myasthenia gravis is found in approximately 1% of patients with Graves disease and, similarly, myasthenia sufferers have an increased incidence of Graves disease.

Respiratory system

As part of the generalized myopathy, respiratory muscle function may be impaired and pulmonary compliance may also be reduced. Dyspnoea may result from these changes and breathlessness secondary to cardiac dysfunction may also be exacerbated. All these changes are reversible with appropriate antithyroid therapy.

Skin and hair

With prolonged elevation of thyroid hormone concentrations, diffuse hair loss is common; the nails are brittle and may become elevated from the nail bed (onycholysis). Palmar erythema may be found; the skin itself feels warm and moist due to increased cutaneous blood flow and sweating.

The skeleton

Prolonged hyperthyroidism may result in significant loss of mineral from the skeleton, which results from increased bone turnover, with resorption of bone exceeding accretion. Hypercalciuria and hyperphosphaturia are also found and urinary hydroxyproline excretion is increased, reflecting the increase in collagen turnover. Although hypercalcaemia is commonly found in hyperthyroid individuals, it is usually mild and resolves with effective antithyroid therapy.

The kidneys: mineral and water balance

Hyperthyroid individuals often complain of increased thirst and mild polyuria, even in the absence of hypercalcaemia or hyperglycaemia, and osmoregulation may be disturbed while biochemical hyperthyroidism persists, though this is rarely clinically significant. Plasma sodium and potassium concentrations do not usually alter, but urinary magnesium excretion increases and plasma magnesium concentrations may be low.

Other endocrine systems

In the female hyperthyroid patient, menstrual irregularities occur sometimes, with scanty menstrual loss and/or irregular cycle length. Although cycles usually remain ovulatory, fertility is significantly reduced. In general, plasma concentrations of free sex steroids are reduced as a consequence of increases in binding proteins, but total sex steroid concentrations may be high. Preferential metabolism of androgens to oestrogens may be responsible for the gynaecomastia seen in a small proportion of hyperthyroid males. Hypothalamo-pituitary responsiveness to exogenous gonadotrophins and basal concentrations of LH and FSH are usually normal.

The turnover of cortisol is increased in hyperthyroidism, but basal concentrations remain normal and the response to physiological stress is preserved in the absence of coincidental adrenal or pituitary disease.

Hyperthyroidism in the elderly

Although weight loss is variable, there may be anorexia leading to a clinical suspicion of malignancy. Patients are often tired and withdrawn with reduced mobility due to proximal myopathy, and the term 'apathetic thyrotoxicosis' has been used to describe this presentation. Atrial fibrillation, which may be associated with cardiac failure, is a common presentation in patients over 70 years, especially in those with multinodular goitre, in whom mild hyperthyroidism may have been present for some years. Any co-existing osteoporosis will be accelerated. Elderly patients often present with a single feature of hyperthyroidism (e.g. weight loss, atrial fibrillation) without the other systemic features more commonly seen in younger patients.

Causes of hyperthyroidism

Table 19.4 lists the causes of hyperthyroidism and summarizes their pathogenesis.

Thyroid stimulating hormone measurement demonstrates suppression of TSH concentrations in all cases of hyperthyroidism (except those due to inappropriate pituitary secretion of TSH, as described later). Free T4 is usually elevated in association with raised free T3 concentrations, though true T3 toxicosis (elevation of T3 alone) may be present. Specific diagnostic biochemical and immunological tests are outlined below.

When following the response of hyperthyroidism to treatment, TSH concentrations should be measured, but as it is clear that TSH suppression may persist for a variable period of time after restoration of normal free T4 and T3 concentrations, the measurement of TSH should not be used unsupported to follow response to therapy. The authors recommend that, ideally, concentrations of free T4 and T3 be measured with TSH until TSH secretion returns to normal.

Graves disease

Graves disease, an autoimmune condition of unknown aetiology, is characterized by the presence of diffuse thyroid enlargement, orbital tissue involvement (Graves

TABLE 19.4 The major causes of hyperthyroidism

Cause	Pathogenesis
Graves disease	Hyperthyroidism due to TSH receptor stimulating antibody
Toxic multinodular goitre	Autonomous function in areas of a nodular gland
Toxic nodule	Autonomously functioning nodule
Chronic iodine excess	High thyroidal iodine ± areas of autonomously functioning gland or thyroid stimulating antibody
Thyroiditis	Transient phenomenon resulting from release of stored thyroid hormone from damaged gland
Pituitary tumour secreting TSH	Inappropriate release of TSH stimulating the thyroid
Amiodarone	High iodine content; alteration of thyroid hormone metabolism
Thyrotoxicosis factitia	Exogenous administration of thyroid hormone
Ectopic thyroid tissue	Functionally active and autonomous thyroid tissue present in extrathyroidal tumours
Trophoblastic tumours	Secretion of thyroid stimulator

ophthalmopathy) and, less commonly, skin involvement (pretibial myxoedema and thyroid acropachy) in addition to thyroid dysfunction. The disease predominantly affects females, with a female:male ratio of approximately 6:1, and peak incidence in the fourth and fifth decades. As with other autoimmune diseases, there is clearly an inherited predisposition to develop the disease, although other family members may have hypo- rather than hyperthyroidism. Excess iodine may induce hyperthyroidism in susceptible individuals, and observational studies indicate a relationship between major stress, such as loss of a close relative, and the onset of hyperthyroidism some months later. It is now well established that stress may modify the immune response.

Graves disease is also seen as the immune system recovers after a period of immunosuppression. This occurs in around 3% of patients with HIV disease after initiation of retroviral therapy and up to 15% of patients with multiple sclerosis 9–12 months after treatment with Campath® monoclonal antibody therapy as the T cell compartment repopulates.

Thyroid involvement. In Graves disease, thyroid enlargement (goitre) is typically moderate and diffuse, and the gland has a soft consistency. Because of increased blood flow through the hyperactive gland, a vascular bruit may be heard over the thyroid.

Eyes. Hyperthyroidism produces a characteristic staring expression, with the sclera being seen above and below the iris (lid retraction) and a tendency for the lid movement to lag behind that of the globe as patients look downward from a position of maximum upward gaze (lid lag). These signs may be present in any patient with hyperthyroidism. Additional and specific orbital signs and symptoms may be found in patients with Graves disease (Fig. 19.8). Similar changes may rarely occur with other



FIGURE 19.8 ■ The appearance of a patient with severe Graves ophthalmopathy. Significant proptosis, periorbital swelling, chemosis and conjunctival infection are all present. The dilated left pupil is a consequence of the use of mydriatic eyedrops and not the disease process. (© University Hospital of Wales, Heath Park, Cardiff Medical Illustration Dept).

TABLE 19.5 The eye signs found in thyrotoxicosis

Sign	Examination finding
Proptosis ^a	Forward protrusion of the globe
Lid retraction	Sclera visible above/below iris
Lid lag	Lid movement lags behind that of globe
Ophthalmoplegia ^a	Failure of full ocular movement often producing diplopia
Chemosis ^a	Oedema of conjunctival sac
Periorbital oedema ^a	Oedema and prolapsed orbital fat present around eye
Conjunctival infection ^a	Inflammatory response
Visible ocular muscle ^a	Enlarged extraocular muscle visible anteriorly
Loss of vision ^a	Compressive optic neuropathy or corneal exposure with scarring
Ptosis (rarely)	Drooping of upper eyelid

^aEye signs specific to Graves disease rather than thyrotoxicosis.

autoimmune thyroid diseases such as Hashimoto thyroiditis and are listed in Table 19.5. Patients with Graves ophthalmopathy should be managed in a multidisciplinary eye clinic with the joint input of ophthalmologists and endocrinologists.

Skin. Patients with Graves disease may sometimes develop indurated purple skin lesions, classically over the anterior tibia. These areas (pretibial myxoedema) contain large amounts of glycosaminoglycans and tend to be seen in patients with high concentrations of thyroid stimulating antibody and ophthalmopathy. Acropachy (similar to clubbing) is a recognized but rare feature of Graves disease.

Diagnosis. The diagnosis of Graves disease is not difficult when the typical triad of hyperthyroidism, goitre and extrathyroidal involvement is present. The diagnosis may be made in other circumstances by demonstrating the presence of thyroid stimulating antibody and is supported by the finding of diffuse, increased uptake of radioisotope on thyroid scintiscanning (see Figure 19.9A).

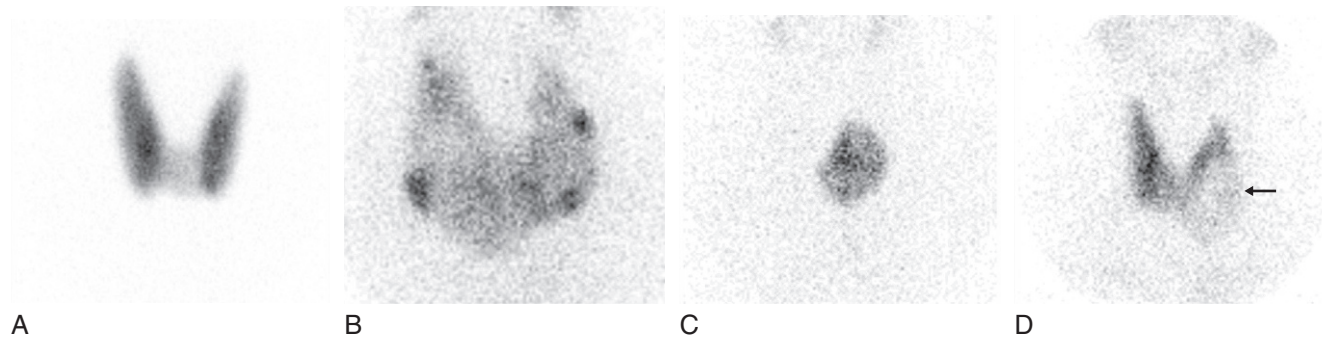


FIGURE 19.9 ■ Isotope scans of patients with various thyroid disorders. The figure shows ^{99m}Tc scan images of (A) Graves disease showing characteristic diffuse uptake of isotope; (B) multinodular goitre with patchy uptake of isotope and inadequate activity in relation to the autonomously functioning nodules; (C) toxic adenoma causing 'T3 thyrotoxicosis'. The suppressed TSH secretion leads to atrophy of the normal thyroid tissue, which recovers following removal of the adenoma by surgery or ^{131}I ablation; (D) a 'cold nodule' with reduced uptake in relation to the rest of the gland. Fine needle aspiration revealed papillary carcinoma confirmed histologically following thyroidectomy.

Natural history. The natural history of the hyperthyroidism of Graves disease is that of successive relapse and remission over many years in the majority of patients. In about 30% there is a single episode of hyperthyroidism, lasting several months, followed by prolonged remission and even the eventual development of hypothyroidism 10–20 years later. Unfortunately, it is not possible at first presentation to identify which patients are destined to have a single episode. The antithyroid drugs, which do not influence the natural history of Graves disease, are given for an empirical period of 9–18 months, thereby identifying the 'single episode' patient and avoiding the use of destructive therapy.

Treatment. There are three forms of treatment of the hyperthyroidism of Graves disease: antithyroid drugs, surgery and ^{131}I . Each is effective but none is perfect. Medical therapy with antithyroid drugs offers the only chance of lasting remission without resulting in permanent hypothyroidism. However, medical treatment is required for at least 9–18 months in most instances and overall success rates are <50%. Surgery or radioactive iodine therapy provide rapid control of hyperthyroidism, but incur the need for lifelong thyroid hormone replacement. Radioiodine treatment with ^{131}I may be associated with a rise in thyroid stimulating antibody concentrations and a worsening of ophthalmopathy and is absolutely contraindicated in pregnancy or the puerperium. Surgery is generally safe and effective, but carries the risks associated with anaesthesia and can be complicated by vocal cord damage or transient or permanent hypoparathyroidism. These complications are, however, minimized in the hands of an experienced surgeon.

The antithyroid drugs carbimazole (and its active form methimazole) and propylthiouracil exert their action principally by inhibiting the action of thyroid peroxidase and, therefore, thyroid hormone synthesis. Treatment is given for 9–18 months. Shorter treatment periods are associated with higher relapse rates, but treatment periods extending beyond 18 months carry no advantage. The antithyroid drugs can be given either in titrated doses or as part of a 'block and replace' regimen using both carbimazole and thyroxine. Both approaches are equally effective but the block and replace regimen, although generally resulting

in more stable thyroid function and a reduced frequency of follow-up visits for dose titration, may be associated with slightly more side-effects than the titration. Lasting remission is only achieved in 30–40% of cases: adverse prognostic indicators include younger age, male gender, high pretreatment concentrations of T3, the presence of a large goitre and persistence of TSH receptor antibody at the end of the planned treatment period. As the drugs do not affect release of thyroid hormone from the gland, there is typically a delay of 4–8 weeks before euthyroidism is restored. Although persistently raised TSH receptor antibody concentrations at the end of therapy are associated with an increased risk of relapse, they are not sufficiently predictive to be of value in guiding clinical decision-making and testing is currently recommended only for diagnostic and not for prognostic purposes.

Symptomatic relief can be obtained in nearly all patients by also using β -blocking drugs for these first few weeks of therapy in order to reduce their tremor and tachycardia. In pregnant thyrotoxic patients, careful monitoring is required in order to keep the dose of antithyroid drugs to a minimum, especially during the last trimester, as these compounds cross the placenta and may render the fetus hypothyroid. 'Block and replace' treatment should not be used as T4 crosses the placenta only poorly. It is generally recommended to maintain free T4 and free T3 concentrations in the high-normal range to ensure sufficient transfer of thyroid hormones to the fetus to counter the effects of transplacental passage of the antithyroid drugs.

Iodine-131 can be used therapeutically to treat hyperthyroidism due to Graves and other thyroid disorders. Radioiodine works initially by interfering with organification of iodine, but later predominantly inhibits replication of thyrocytes, by inducing radiation damage in the gland and hence controlling thyroid overactivity, after a delay of weeks or months. The major side-effect of this form of therapy is long-term hypothyroidism, with up to 80% of patients becoming hypothyroid within the first year, if a standard dose of 400 MBq is used. In some individuals, there is a transient rise in thyroid hormone concentrations in the first few weeks after radioiodine before they begin to fall. In patients with Graves disease, the appearance or worsening of thyroid

eye disease may occur, and in patients with active eye disease it is recommended that radioiodine is avoided and/or steroid prophylaxis for 6–8 weeks is given (e.g. prednisolone 20 mg reducing by 5 mg every two weeks). Graves ophthalmopathy is associated with smoking and cessation should be advised, particularly prior to radioiodine treatment. There is no evidence of an increase in the risk of malignancy following therapeutic doses of radioiodine. The dose should be repeated if hyperthyroidism persists after six months. Note that transient hypothyroidism at three months is occasionally seen with return of thyrotoxicosis by six months following radioiodine requiring repeat therapy.

Thyroidectomy is a highly effective treatment for Graves disease. Patients should be rendered euthyroid prior to operation using antithyroid drugs. In some centres, the drug is stopped for two weeks prior to surgery and potassium iodide substituted; this reduces the size and vascularity of the gland and makes surgery technically easier. To avoid the risk of relapse and the requirement for repeat surgery, near-total thyroidectomy (rather than subtotal thyroidectomy) is now routinely performed for thyrotoxicosis, rendering all patients hypothyroid post-operatively, and treatment with thyroid hormone is routinely administered immediately post surgery.

Treatment of Graves ophthalmopathy. A third of patients with Graves disease have thyroid eye involvement; sight loss is rare but may occur if untreated in 3–5% of patients. In addition, most patients have impaired quality of life and suffer significant psychological distress due to the disfiguring appearance of the eyes. In its classic form thyroid eye disease is easily recognizable. However, some patients suffer with eye disease for months to years before a diagnosis, which may be missed in patients without eye symptoms, patients with unilateral eye disease or patients who are euthyroid at presentation since orbitopathy may precede or post-date the development of hyperthyroidism. General management measures include smoking cessation and maintenance of normal thyroid status. Radioiodine therapy is best avoided in patients with moderate to severe eye disease (see above) but may be administered with glucocorticoid cover in mild cases. Local measures, e.g. lubricants, ointments and corrective prisms, will be adequate for mild disease. Glucocorticoids are first-line therapy in active moderate to severe disease. Other treatments that have been tried include ciclosporin, methotrexate, intravenous immunoglobulins, orbital radiotherapy, and the B cell depleting agent, rituximab. Selenium is an antioxidant with a non-specific anti-inflammatory effect that has been shown to reduce symptoms in mild Graves ophthalmopathy.

Toxic multinodular goitre

Hyperthyroidism arising in a previously multinodular goitre occurs in an older population than that affected by Graves disease; patients are typically over the age of 50, with females being affected more than males.

Clinical features. The cardiovascular features of hyperthyroidism tend to be prominent in this, often elderly,

group of patients, though all the features of hyperthyroidism mentioned earlier may occur. The goitre itself is classically nodular and may be large, often having been present for many years prior to the onset of thyroid dysfunction.

Diagnosis. The biochemical diagnosis of hyperthyroidism in this situation is fairly straightforward, with suppression of TSH, though thyroid hormone concentrations may not be grossly elevated, in some cases lying at the upper limit of normal. Thyroid autoantibodies are not usually present. Thyroid scintiscanning shows a patchy uptake of isotope, with multiple hot and cold areas being seen throughout the gland (Fig. 19.9B). Radioiodine uptake values may lie at or above the upper limit of normal, but are not usually grossly elevated, and thyroid ultrasound can be used to confirm the multinodular nature of the gland.

Treatment. In general, radioiodine is the treatment of choice for the patient with toxic multinodular goitre as relapse invariably occurs after withdrawal of antithyroid drug therapy. Antithyroid drugs can be used until radioiodine becomes effective. Where compressive symptoms result from gland enlargement, surgery should be used in place of radioiodine.

Toxic adenoma

Isolated, autonomously functioning adenomas within the thyroid account for 5% of cases of hyperthyroidism. Half of these patients have 'T3-thyrotoxicosis' and the resulting clinical hyperthyroidism is usually mild.

Diagnosis. Scintiscans of the gland in cases of toxic adenoma show an area of high uptake in the region of the nodule, with suppression of uptake in the remaining, normally responsive, areas of the gland (Fig. 19.9C).

Treatment. Patients with proven toxic adenoma may be treated successfully either with radioiodine, which is concentrated by the hyper-functioning adenoma, or by surgery. Radioiodine is usually the treatment of choice and post ¹³¹I hypothyroidism is less common than with Graves disease.

Thyroid stimulating hormone-secreting pituitary tumour

Very rarely, adenomas of the pituitary gland secreting TSH (TSHomas) may produce hyperthyroidism.

Diagnosis. The biochemical key to making this diagnosis is the persistence of TSH secretion despite definite overproduction of thyroid hormones. The differential diagnoses are thyroid hormone resistance (see p. 400), drugs (amiodarone/amphetamine), and acute psychotic illness. Interference with one or other of the thyroid function assays should always be excluded, for example by repeating the tests on another assay platform. In TSHomas, high circulating concentrations of the α -subunit of TSH may be found in the serum, TSH secretion is not increased by

the administration of TRH and a goitre is present in the absence of immunological evidence of thyroid autoimmunity. Pituitary MRI or CT imaging should be performed, but not before the biochemical diagnosis is established, as the lesion may be too small to visualize and because incidental non-functioning microadenomas of the pituitary are common. If large, the pituitary lesion may produce local damage, with reduction or loss of secretion of other anterior pituitary hormone and impairment of the visual fields. TSH secreting pituitary adenomas may co-secrete other anterior pituitary products such as prolactin and growth hormone.

Treatment. Treatment consists of surgical removal of the tumour or, alternatively, pharmacological reduction of TSH secretion. The long-acting somatostatin analogue octreotide has been reported to reduce TSH secretion and tumour size in individual cases and dopaminergic agonists such as bromocriptine and cabergoline may also be effective. Ablative antithyroid therapy will lead to control of hyperthyroidism, but will not deal with the primary problem and is, therefore, not generally indicated.

Other causes of hyperthyroidism

In some of the areas discussed below, the term 'hyperthyroidism' is strictly not applicable, as the elevation of thyroid hormones derives from extrathyroidal sources. However, for continuity, this term will be used to represent states of clinical and biochemical thyroid hormone excess.

Iodine. In individuals with goitre due to previous iodine deficiency, chronic administration of excess iodine in the diet can induce a hyperthyroid state. This phenomenon (sometimes called the Jod-Basedow phenomenon) is particularly likely to occur in patients who already have a degree of pre-existing thyroid autonomy, expression of which may have been masked by the lack of iodine. In iodine-replete areas, a similar phenomenon may be seen, especially in individuals with a multinodular goitre and in some cases the hyperthyroidism reverses completely when the source of additional iodine is withdrawn from the diet. The precise relationship between iodine dose and thyroid response is clearly complex, since it is dependent not only on the individual's prior exposure to iodine and the degree of gland autonomy, but also on the time course of iodine administration; use of large doses of iodine as a therapeutic means of rapidly treating hyperthyroidism has quite the opposite of the effect seen with chronic administration of more moderate doses. In iodine-induced hyperthyroidism, radioiodine uptake is characteristically reduced and urinary iodine excretion increased, with coincident TSH suppression and evidence of elevated T4 and elevated or high normal T3 concentrations.

Amiodarone. This anti-arrhythmic drug contains iodine and, at therapeutic dosage, generates 6–12 mg of free iodine per day compared with the recommended iodine intake of 0.2 mg/day. The drug can produce abnormalities in thyroid function tests in up to 50% of patients, but may also induce true thyroid dysfunction. Amiodarone has a

structure similar to that of thyroid hormones and inhibits the peripheral conversion of T4 to T3 catalysed by iodothyronine deiodinase D1. Resulting concentrations of T4 may, therefore, be high and T3 low, with an increase in rT3. Thyroid stimulating hormone may rise transiently during the first few weeks of treatment, but by four months, most patients who are euthyroid will have normal or sometimes suppressed concentrations of TSH. The drug inhibits both iodine uptake by the thyroid and entry of T4 into cells and can also cause both iodine-induced hypothyroidism and hyperthyroidism. The frequency of these two conditions depends on the ambient iodine intake. In areas of high iodine intake (e.g. Worcester, USA), amiodarone produces hypothyroidism in up to 20% of patients receiving the drug, while hyperthyroidism only occurs in about 2% of patients. Conversely, in areas of Italy, where iodine intake is low, hypothyroidism occurs in 5% of patients and hyperthyroidism in 10%. In patients with an underlying thyroid abnormality (e.g. multinodular goitre or subclinical Graves disease), the high iodine content can induce thyrotoxicosis (amiodarone-induced thyrotoxicosis type I) that is treatable with antithyroid drugs), but the drug may also induce a destructive thyroiditis (amiodarone-induced thyrotoxicosis type II) that is not readily responsive to antithyroid drugs, but is often responsive to steroids. A mixed picture is also seen, and occasionally urgent thyroidectomy is required for thyrotoxicosis refractory to both forms of therapy. It is therefore important to evaluate patients before they commence therapy with amiodarone. This should include clinical examination and a basal measurement of TSH, free T4, free T3 and TPOAb. After starting treatment, thyroid function tests should be repeated at six months and thereafter every six months, including the year after the drug is stopped. The long half-life of the drug means that changes in thyroid test results may persist for some time after ceasing therapy.

Thyrotoxicosis factitia. Cases of hyperthyroidism occur due to self-administration of thyroid hormone, taken either as T3 or T4. Radioiodine uptake is low, goitre is absent and thyroglobulin concentrations are low. Unless there is a prior history of thyroid autoimmunity, thyroid-directed antibodies will be absent. Thyroid stimulating hormone will be undetectable; total and free T4 concentrations will be high if the patient is taking T4 and suppressed if the patient is taking T3, and total and free T3 concentrations will be high in both cases. Careful supervision of patients with serial measurement of thyroid function in hospital may be necessary, and specific psychological assessment and counselling may be required after the diagnosis has been made.

Ectopic thyroid tissue. Metastatic thyroid follicular carcinoma may rarely produce sufficient thyroid hormone to result in hyperthyroidism. Other tumours, such as ovarian teratomata, may contain functional thyroid tissue in sufficient quantity to produce symptoms and signs of thyrotoxicosis (struma ovarii): endogenous thyroid radioisotope uptake will be suppressed and functioning tissue demonstrable in the tumour under these most uncommon circumstances.

BOX 19.3 Indications for treatment in patients with subclinical hyperthyroidism

- Cardiac arrhythmias
- Congestive cardiac failure
- Osteoporosis
- Presence of hyperthyroid symptoms
- Goitre
- Pregnancy
- Infertility

Other thyroid stimulators. Trophoblastic tumours such as choriocarcinoma, hydatidiform mole and metastatic embryonal testicular carcinoma may secrete human chorionic gonadotrophin (hCG) and produce hyperthyroidism.

Subclinical hyperthyroidism. This term is used to describe patients with the biochemical picture of suppressed serum TSH but normal concentrations of T3 and T4. Subclinical hyperthyroidism is present in about 0.5–3.0% of the population and rises in prevalence with age. Most patients are asymptomatic and revert to normal thyroid status over time. However, a small proportion of patients (1–5% per annum) will progress to overt hyperthyroidism (suppressed TSH and elevated T3 and T4). Patients with a fully suppressed TSH concentration (<0.01 mU/L) are more likely to progress to overt hyperthyroidism than those with partial TSH suppression (TSH 0.1–0.4 mU/L). Patients with subclinical hyperthyroidism are at increased risk of atrial fibrillation and osteoporosis and, once persistent abnormal thyroid function over a period of six months is confirmed, treatment may be indicated to prevent these complications. Other indications for treatment are listed in [Box 19.3](#). Radioiodine is generally used as the underlying aetiology is almost always nodular thyroid disease, although this form of treatment is contraindicated in pregnancy.

Hyperthyroidism or non-thyroidal illness?

It may be difficult to determine whether the finding of suppressed serum TSH and raised free T4 are due to non-thyroidal illness (NTI) or hyperthyroidism, especially if the serum T3 concentration is in the upper part of its reference range. A raised total T4 is uncommon in NTI. Furthermore, features such as ophthalmopathy, goitre or unexplained atrial fibrillation will favour hyperthyroidism. In the absence of these clinical signs, measurement of the TSH receptor antibody and isotope imaging may be helpful in detecting Graves disease and nodular goitre. If there still remains doubt, repeat thyroid function testing after 3–6 months is indicated.

HYPOTHYROIDISM

Clinical features

The clinical features seen as a consequence of low circulating concentrations of thyroid hormone are summarized in [Table 19.6](#).

TABLE 19.6 The symptoms and signs of hypothyroidism

	Common	Uncommon
Symptoms	Lethargy Skin dry and coarse Slow speech and mentation Facial puffiness (oedema) Cold intolerance Pallor Hoarse voice Constipation Weight gain Hair dry and falling out Weakness/stiffness Breathlessness	Anorexia Irritability Menorrhagia Angina Deafness Poor coordination
Signs	Periorbital/facial oedema Pale, dry skin Goitre Obtunded mental state Cool peripheries Diffuse alopecia Bradycardia Median nerve compression Delayed relaxation phase of reflexes	Psychosis Pleural/pericardial effusions Ascites Cerebellar ataxia Galactorrhoea Hydrocoele (males)

Cardiovascular system

A reduction in resting cardiac output occurs. Cool peripheries are characteristic, owing to a reduction in cutaneous blood flow, which contributes to the intolerance to cold that is so often present. In the thorax, cardiac dilatation may be seen on chest X-ray and occurs in association with pericardial effusion, which occasionally leads to compromised myocardial function. Ischaemic chest pain is uncommon, more frequently being seen in hypothyroid patients with coincidental ischaemic heart disease when they first receive thyroid hormone replacement therapy. The cardiac changes reverse with appropriate thyroid replacement therapy.

Gastrointestinal system

Most individuals show a moderate weight gain despite reduced appetite, due primarily to fluid retention. Intestinal absorption of nutrients may be affected both by reduced absorption rates and increased intestinal transit time, but net absorption may actually be reduced, normal or even increased. Results of biochemical liver function tests are usually normal.

Central and peripheral nervous system

Deficiency of thyroid hormones in fetal or early neonatal life, if not promptly treated, results in irreversible damage to the CNS, with structural abnormalities evident on histological examination. The classic picture of congenital hypothyroidism is fortunately only rarely seen now,

due largely to the introduction of neonatal screening for hypothyroidism and the efficacy of therapy, if introduced sufficiently early in life.

In adult life, neurological defects resulting from hypothyroidism are usually reversible. The characteristic features are of generalized slowing in intellectual function, with inanition, slow mentation, somnolence, and occasionally, a frankly psychotic state. Speech becomes slow and the voice coarse and gruff in nature, the latter in part due to oedema within the vocal apparatus. Cerebellar ataxia may be seen with prolonged hypothyroidism and may become irreversible with delay in treatment. Seizures may also occur in severe cases. Peripheral nervous system manifestations are also common, with compression of the median nerve at the wrist being perhaps the best known (carpal tunnel syndrome). Relaxation of the tendon jerks is characteristically delayed.

Locomotor system

Muscular stiffness is a particularly common complaint in hypothyroidism and relates to reduced relaxation rate. The muscles show abnormal structure on microscopy, with loss of striations, oedema, swelling of fibres and relative deficiency of type II fibres. Muscular weakness is often evident clinically and plasma activities of muscle enzymes such as creatine kinase may be raised.

Respiratory system

Similar changes to those described above may occur in the respiratory muscles, with abnormal muscle function leading, in patients with pre-existing lung disease, to exacerbation of any carbon dioxide retention and sleep apnoea. Chest X-ray may show pleural effusions, though these are seldom large.

Skin and hair

Increased water binding occurs as a result of deposition of mucopolysaccharides in the skin, in common with the other tissues. The indurated oedema that results gives rise to the 'myxoedematous' appearance of the typical hypothyroid patient. Associated anaemia and hypercarotenaemia through impaired conversion of β -carotene into retinol may render the skin pale or yellow, respectively. Body hair tends to be easily lost, though the classic description of loss of the outer-third of the eyebrows is rarely seen today.

The skeleton

Thyroid hormone deficiency in early life leads to abnormalities of the epiphyses with marked reduction in linear growth and stunted final height. Short stature may be an apparently isolated presentation of hypothyroidism in childhood. In common with all systemic illnesses, prolonged hypothyroidism in childhood leads to retardation of bone age compared with chronological age. Rates of bone turnover are reduced, leading to a reduction in the pool of exchangeable calcium. Plasma

concentrations of calcium and phosphate remain normal. Alkaline phosphatase activity tends to be low in children with hypothyroidism.

The kidneys: mineral and water balance

Renal blood flow and glomerular filtration are both decreased, but total body water has been shown to increase with hypothyroidism, owing to impaired renal excretion of water that itself results from a reduction in delivery to the distal tubule and abnormal osmoregulatory function in the hypothalamus and posterior pituitary. Although exchangeable body sodium is increased, the dilutional effect typically leads to a mild hyponatraemia. Plasma creatinine and urea concentrations remain normal.

Reproductive system

In adults of both sexes, hypothyroidism leads to a reduction in libido and subfertility. Menorrhagia due to failed progesterone secretion with anovulatory cycles is common in the female, as is oligospermia in the male. These changes may be related to impaired luteinizing hormone secretion, particularly in longstanding cases. Basal gonadotrophin concentrations are, however, typically within the normal range, unless pituitary disease is responsible for the hypothyroid state. A reduction in sex hormone binding globulin concentration leads to an elevation in free sex hormone concentrations, with a reduction in total concentrations of both oestrogen and testosterone following from both this and alterations in sex steroid synthesis.

Other systems

Cortisol turnover is often reduced and, in some patients, the cortisol response to hypoglycaemia is blunted, although response to exogenously administered ACTH is normal. Hyperprolactinaemia is a common finding, resulting from increased TRH release. The hyperprolactinaemia correlates with the degree of elevation of TSH, although frank galactorrhoea is only relatively infrequently found.

A normochromic anaemia, which may be either normo- or macrocytic in nature, is often seen and is likely to reflect diminished production of erythropoietin, with low red cell mass. Macrocytosis is frequently ascribed to coincidental occurrence of vitamin B₁₂ deficiency due to autoimmune pernicious anaemia, but a deficiency in the response to available B₁₂ and concurrent malabsorption of folate from the gut may contribute. If menorrhagia has been prolonged or severe, microcytosis, rather than macrocytosis, may be present as part of an iron deficiency state. Plasma concentrations of clotting factors VIII and IX may be reduced in hypothyroidism and platelet adhesiveness reduced, causing a mild bleeding tendency that may exacerbate the existing risk of anaemia. The effects of thyroid hormones on lipids have been outlined earlier in this chapter. Glucose absorption from the gut and uptake by the tissues from plasma are both delayed, but in patients with established diabetes, a characteristic increase in sensitivity to exogenous insulin is seen, probably due to decreased clearance of insulin.

Causes of hypothyroidism

Causes of hypothyroidism are listed in [Table 19.7](#).

Measurement of plasma TSH concentrations provides the cornerstone of the biochemical evaluation of hypothyroidism. As circulating concentrations of thyroid hormone fall, TSH secretion increases and is used to monitor thyroid status. Secretion of T3 is preferentially maintained in the presence of the high TSH concentrations that accompany declining thyroid function. T3/T4 ratios, therefore, rise and plasma T4 concentrations correlate better with thyroid activity than do plasma T3 concentrations. Replacement therapy for primary hypothyroidism can be monitored using plasma TSH concentrations. The particular problems resulting from pituitary failure are described below. Hashimoto disease and other forms of thyroiditis are described in the following section on thyroiditis. It should be noted that not all raised concentrations of TSH equate to hypothyroidism: of possible sources of error clinically, the most important is hypoadrenalism, which results in a mild to moderately raised TSH (with completely normal free T4) but treatment with thyroxine (rather than steroids) will increase metabolism of any remaining cortisol and precipitate a potentially fatal hypoadrenal crisis.

Primary myxoedema

In a large proportion of patients with primary hypothyroidism, no goitre, or history of goitre, will be present. These individuals will have high plasma TSH concentrations with low total and free T4 concentrations, and the majority will have anti-peroxidase and/or anti-thyroglobulin antibodies. Thyroid growth blocking antibodies have also been described in these patients. While the majority of patients may represent end-stage Hashimoto thyroiditis with lack of clinical recognition of the early phases of the disease, clearly other immunological processes may be responsible for some of them.

Post-surgery or post-radioiodine

Although the hypothyroidism developing after surgery or ¹³¹I therapy for Graves disease is usually permanent, it may be temporary, especially in the first 3–4 months. If symptomatic at this early stage, patients should be treated with a relatively small dose of thyroxine (50–75 µg daily) and the need for continued treatment assessed at six months. If serum TSH remains elevated, thyroxine therapy will be required lifelong and the dose needs to be increased. If at that stage serum TSH is normal or undetectable, stopping the thyroxine for four weeks and re-testing thyroid function would be the appropriate action.

Congenital hypothyroidism

Causes of congenital hypothyroidism are listed in [Box 19.4](#). Clinical detection of hypothyroidism in the neonate may be difficult. In many developed countries, screening programmes measuring TSH or T4 on dried bloodspots from heel-prick blood samples (taken at six days when phenylketonuria screening is performed) have been operating for some years with considerable success. Abnormal thyroid development is responsible for most congenital hypothyroidism. Scintiscanning may be used in order to help to clarify the diagnosis.

Inability to organify trapped thyroidal iodide can be recognized by an abnormal perchlorate discharge test and is usually associated with goitre. Patients with mild forms of this type of defect may also have deafness (Pendred syndrome), but are in this case not often hypothyroid. Ineffective transport of iodide into the thyroid gland can produce both goitre and hypothyroidism and will be found in individuals with abnormally low radioiodine uptake. The block to transport can, in some cases, be overcome by treating patients with high doses of oral iodide, reducing goitre and restoring normal thyroid function.

Deficiency of iodotyrosine dehalogenase produces impaired deiodination of iodotyrosine, such that large

TABLE 19.7 Causes of hypothyroidism

	Disease	Pathogenesis
Primary (due to disease of the thyroid gland)	Thyroiditis: Hashimoto; postpartum; de Quervain (rarely); Riedel (rarely)	Autoimmune destruction of the gland Prolonged inflammatory damage Gland becomes fibrosed
	Primary myxoedema	Gland atrophy; post-autoimmune in many cases
	Ablative	Post-radioiodine or surgery; gland bulk reduced
	Athyreosis	Abnormal or absent thyroid tissue
	Dyshormonogenesis	Abnormal synthesis of thyroid hormones
	Drugs	Inhibition of thyroid hormone synthesis or induction of thyroid autoimmunity
	Iodine deficiency	Lowered gland iodine; impaired hormone synthesis
	Acute iodine excess	Transient inhibition of hormone synthesis may become permanent in the presence of coexisting thyroid destructive autoimmune activity
Secondary (due to extrathyroidal disease)	Anterior pituitary failure: tumours; granulomatous deposits; Sheehan syndrome; post pituitary destructive therapy; developmental abnormalities; post-head injury	Loss of TSH stimulation of thyroid
	Hypothalamic dysfunction	Loss of TRH stimulation of anterior pituitary

BOX 19.4 Causes of congenital hypothyroidism

- Structural abnormalities
 - Absent gland
 - Ectopic site
- Enzyme defects (dyshormonogenesis)
 - Iodide transport
 - Iodide organification
 - Iodotyrosine dehalogenase
- Defective iodoprotein secretion
- Mutation in MCT8 gene (T3 transporter)
- Maternal antibodies (transient)

amounts of mono- and di-iodotyrosines appear in the blood, and iodine deficiency can result from the urinary loss of large quantities of deaminated metabolites of these molecules. Again, large doses of iodine may reduce the severity of the hypothyroidism and the accompanying goitre. Biochemically, parenteral administration of mono- and di-iodotyrosines will be followed by their rapid urinary excretion and this can be used as a diagnostic test for the condition.

Diversion of thyroidal iodine into metabolically inactive or less active iodoproteins, rather than thyroid hormones, can lead to hypothyroidism. Radioactive iodine uptake will be high, due to high TSH drive to synthetic activity within the gland, and the abnormal quantities of iodinated products of the gland can be measured in serum. Thyroid hormone replacement therapy is required.

Maternal TSH receptor-blocking antibodies can cross the placenta to produce transient hypothyroidism in the fetus, which resolves within a few weeks after birth. In these patients, bloodspot TSH values are usually only modestly raised. Testing for TSH receptor antibodies would distinguish these infants from those requiring T4 replacement therapy.

Lithium treatment

Lithium therapy may lead to reduced thyroid hormone synthesis in a similar manner to the acute effects of iodine, but more commonly aggravates any underlying autoimmune disease. Female patients are more often affected and thyroid autoantibodies are often present. Lithium can, therefore, cause hypothyroidism (in approximately 14% of women and 5% of men taking lithium) but also, less commonly, induces hyperthyroidism. Patients taking lithium should have their thyroid function tests measured every six months. Patients on lithium who have positive TPO antibodies are at increased risk of developing thyroid dysfunction. Lithium itself has been used as a treatment for hyperthyroidism and as an adjunct to radioiodine treatment (as it enhances radioiodine uptake) but because of the potential for toxicity, this approach has not entered routine practice.

Cytokine therapy

Treatment with a variety of cytokines (see Table 19.1) can result in hypothyroidism that is more common in patients with underlying autoimmune thyroid disease. This

is most commonly observed in patients being treated with interferon therapy for viral hepatitis. It is usually self-limiting once cytokine therapy ceases, but may need temporary treatment with thyroxine.

Iodine

Both excess and deficiency of dietary iodine may result in hypothyroidism.

Although oral use of large doses of iodine would normally lead to transient suppression of thyroid hormone synthesis, iodine can rarely produce more lasting hypothyroidism, particularly in individuals with Hashimoto disease.

Environmental iodine deficiency remains a major cause, worldwide, of goitre and hypothyroidism. Clinical euthyroidism is often maintained by preferential production of T3 and compensatory thyroid hyperfunction, leading to large goitre formation.

Secondary hypothyroidism

Loss of TSH drive to thyroid function may result from any of the causes of pituitary gland hypofunction. Recognition of the aetiology of this type of hypothyroidism is critical and is usually made by observing a low T4 accompanied by an inappropriately 'normal' TSH concentration. In some cases of secondary hypothyroidism, TSH may be slightly increased. Increased T3:T4 ratios are not seen in this type of hypothyroidism as TSH drive is absent.

Treatment of hypothyroidism

Myxoedema coma

Profound hypothyroidism may produce the state known as myxoedema coma. This is a relatively uncommon endocrine emergency that may be triggered by factors such as infection, cold exposure, the use of sedative medications or non-compliance with treatment. The classic presentation is that of decompensated hypothyroidism with altered sensorium, marked bradycardia, hypothermia and circulatory collapse. Patients should be admitted to an intensive care unit where possible. Therapy is supportive with thyroid hormone replacement using nasogastric T4 or parenteral T3 therapy. Replacement doses of glucocorticoids should be given until coincidental adrenal insufficiency has been excluded, infections should be treated aggressively and further heat loss prevented to allow progressive rewarming to occur. Despite appropriate therapy, the mortality from this condition remains high, often with arrhythmias in the first 24 h after commencing therapy, and early recognition and careful management in a critical care setting offers the best chance of recovery.

Thyroid hormone replacement therapy

Treatment of primary or secondary hypothyroidism is with oral synthetic levothyroxine. The usual replacement dose is 1.6 µg/kg per day, which equates to about 100–125 µg daily. Patients are customarily advised to take levothyroxine 30–60 min before breakfast since food and caffeine interfere with its absorption. Bed time administration is

equally effective and may be suited to patients who take a number of other medications. Treatment can usually be initiated at full replacement doses except in elderly individuals and patients with cardiac disease who will require more cautious starting doses. Serum T4 and TSH should be checked 6–8 weeks after initiation, making any minor adjustment to the dose necessary to restore TSH to normal. In the elderly and in patients with established ischaemic heart disease, levothyroxine replacement should begin with 25 µg daily, increasing by 25 µg increments every 4–6 weeks. In patients with hypopituitarism, with both secondary adrenal and thyroid failure, it is important that levothyroxine is not started before hydrocortisone replacement, as this might precipitate adrenal crisis by increasing the metabolic rate in a patient unable to secrete cortisol.

The objective of treatment is to achieve clinical well-being and restore biochemical euthyroidism, i.e. a serum TSH within the laboratory reference range. Patients who remain symptomatic despite full replacement doses may benefit from additional dose increases to keep TSH concentration in the lower half of the reference range. Lowering TSH to below the reference range is not advisable as there is evidence that such individuals with exogenous hyperthyroidism have an increased risk of atrial fibrillation and osteoporosis. The response to treatment in patients with secondary hypothyroidism should be based on FT4 measurements since TSH secretion is deficient in pituitary disease. The therapeutic goal in such cases would be a FT4 in the upper-third of the reference range.

Although thyroid hormone replacement with synthetic levothyroxine is effective in most cases, a significant proportion of patients report that they remain symptomatic. The management of such patients can be challenging and factors that may predispose to inadequate thyroid hormone replacement include poor patient compliance, inadequate dosage and administration, concurrent use of medications which interfere with levothyroxine availability and the presence of comorbidities which impair levothyroxine absorption. The combination of a raised serum T4 and TSH is suggestive of poor compliance, the patient having been over zealous in the few days prior to thyroid function testing in an attempt to compensate for missing doses earlier. Some individuals may remain symptomatic even after normalization of thyroid function. Although some of such patients may benefit from further dose titrations to attain a low normal TSH, it is important to consider alternative causes of persistent symptoms such as diabetes mellitus, obesity, Addison disease, anaemia, depression, coeliac disease, obstructive sleep apnoea and post-viral syndromes. In practice, such investigations are frequently negative and whether some individuals with persistent symptoms will benefit from combined treatment with T4 and T3 remains contentious. The use of T3 in replacement is much harder to titrate and monitor, is not recommended by professional bodies and should not be conducted away from specialist centres with programmes of long-term follow-up for these individuals. Recent delineations of common genetic variations in the deiodinase and thyroid hormone transport proteins suggest that such polymorphisms may account for some of the variability in the response to thyroid hormone treatment.

BOX 19.5 Indications for treatment in patients with subclinical hypothyroidism

- Increased TPO antibody concentration
- TSH >10.0 mU/L
- Goitre
- Pregnancy
- Infertility
- Hypothyroid symptoms^a
- Diabetes mellitus^b
- Cardiac disease^b
- Dyslipidaemia^b

^aSome practitioners advocate treatment of subclinical hypothyroidism if there are classic hypothyroidism symptoms, but this is controversial and the evidence base is lacking.

^bConsider treatment because of increased cardiovascular disease risk.

Subclinical hypothyroidism

This is the term used to describe patients who are clinically euthyroid with normal serum T4 concentration but raised TSH, usually <10 mU/L. There may be mild non-specific symptoms, such as tiredness, weight gain and depression. Now viewed as the mildest form of thyroid failure, it is common practice to treat with thyroxine if antibodies are positive and/or serum TSH is >10 mU/L in order to 'nip things in the bud' before the progression to overt thyroid failure (5% per year), and the possibility of loss to follow-up and later presentation with severe hypothyroidism. Other potential indications for treatment in patients with subclinical hypothyroidism are shown in Box 19.5. Recovery from non-thyroidal illness should be excluded before a diagnosis of subclinical hypothyroidism is made, emphasizing that in all cases thyroid function tests should be repeated and confirmed as remaining abnormal before initiating therapy.

THYROIDITIS

Inflammation of the thyroid gland (thyroiditis) can result in both overactivity and underactivity of the gland.

Thyroiditis producing hyperthyroidism

Autoimmune thyroiditis (typified by Hashimoto disease) is a common cause of hypothyroidism and, as such, will be reviewed in detail below. However, in the early phase of autoimmune damage to the gland, release of stored thyroid hormone may lead to transient hyperthyroidism in some patients. This usually lasts about 6–8 weeks, followed by mild and equally transient hypothyroidism and then recovery, although, in the long term, permanent hypothyroidism is common. Known as silent thyroiditis, the biochemical pattern is similar to that observed in postpartum thyroiditis and in de Quervain thyroiditis.

Diagnosis and treatment

Depending on the phase of the illness, thyroid function may show hyperthyroidism, euthyroidism or hypothyroidism. Radioactive iodine scans in the hyperthyroid phase of the illness will typically show reduced isotope

uptake since the hyperthyroidism is not due to excessive synthesis of thyroid hormones but to the release of pre-formed hormones. For the same reason, treatment with antithyroid drugs is ineffective, although a β -blocker may help to alleviate symptoms of hyperthyroidism.

Hypothyroidism resulting from Hashimoto thyroiditis

In Hashimoto disease, the commonest form of autoimmune hypothyroidism, destructive thyroiditis results from both a cell-mediated and humoral attack on the thyroid tissue. Females are affected with greater frequency than males. The gland is typically enlarged but small and firm in early cases, with a palpable pyramidal lobe. At this stage, hyperthyroidism may transiently occur, but further gland damage rapidly leads to permanent hypothyroidism, at which stage goitre regresses and the gland remnant is composed of fibrous tissue.

Diagnosis

Thyroid-directed autoantibodies (described earlier) are almost invariably present in the early phase of the disease, though they may ultimately become undetectable over a period of years. In about half of patients with Hashimoto thyroiditis, serum TSH is elevated and T4 normal (sub-clinical hypothyroidism), the other patients having overt thyroid failure.

Immunological and biochemical evidence of other organ-specific autoimmune diseases such as diabetes, pernicious anaemia, Addison disease and hypoparathyroidism may also be found.

Other forms of thyroiditis

De Quervain thyroiditis is an unusual condition, which often follows a viral illness. Patients typically present with pain in the thyroid and neck region with fever and malaise. Thyroid function tests may transiently become hyperthyroid, then hypothyroid, before returning to normal. Rarely, permanent hypothyroidism may follow repeated episodes. Riedel thyroiditis is a condition of unknown aetiology. Extensive replacement of the thyroid with fibrous tissue causes the gland to become stony hard. Hypothyroidism is a rare complication.

Hypothyroidism and the postpartum period

Postpartum thyroiditis, due to autoimmune thyroid disease, may develop during the early postpartum period, and is typically histologically and biochemically identical with Hashimoto disease. The classic presentation is of a transient hyperthyroid phase which is followed by a hypothyroid phase. Published series suggest that the overall incidence of the condition is between 5% and 9% and that 40% of patients have failed to recover euthyroid function by one year post-delivery. Of those who do recover normal thyroid function, a further 20–30% will have become hypothyroid within 3–4 years. After the initial episode, the future risk of developing the syndrome increases in successive pregnancies.

Although the presence of antithyroid antibodies during pregnancy is a risk factor for postpartum thyroid disease, at most only half of these women will go on to develop frank thyroid dysfunction. Swings in thyroid function may contribute to the aetiology of postnatal depression and thyroid function should be tested in all such patients.

NEOPLASIA

Thyroid neoplasia forms a small but clearly important proportion of all thyroid disease. Thyroid cancer is the most common endocrine malignancy and accounts for more mortality than all the other endocrine cancers combined. The types of primary and secondary thyroid neoplasm are summarized in [Box 19.6](#). The differentiated thyroid cancers (DTC), papillary and follicular thyroid cancer, are the most common histological types. Papillary thyroid cancer is most frequent in iodine sufficient parts of the world and is associated with a favourable prognosis. Papillary microcarcinomas (diameter <1 cm), are now increasingly detected owing to the growing use of imaging techniques such as ultrasound, computerized tomography or positron emission tomography scans. Anaplastic thyroid cancer is relatively uncommon and typically follows an aggressive course with rapid thyroid enlargement and poor prognosis. Primary thyroid lymphomas almost always arise on a background of Hashimoto thyroiditis and also present with rapid gland enlargement.

Diagnosis

The usual presentation is with a solitary nodule in the thyroid. Although thyroid nodules are common in the general population, only about 5% of such nodules are malignant. Malignancy is suspected in patients with thyroid nodules if there is a history of rapid enlargement, a family history of thyroid cancer or a past medical history of head and neck irradiation as therapy for another disease ([Box 19.7](#)). Clinical suspicion of the diagnosis is further raised by the finding of a nonfunctioning solitary nodule on thyroid scintiscanning ([Fig. 19.9D](#)) that is shown on ultrasound examination to be solid (although cystic lesions may also be malignant). The diagnosis of thyroid carcinoma ultimately depends on histological examination of the gland following an excision biopsy or a fine needle aspirate of the nodule. Biochemical assessment (other than establishing the usually euthyroid status of the patient and excluding evidence of autoimmune processes) plays little part in making the diagnosis. Measurement of

BOX 19.6 Typical thyroid malignancies

Primary neoplasms

- Papillary carcinoma
- Follicular carcinoma
- Anaplastic carcinoma
- Medullary carcinoma (calcitonin secreting; may form part of the syndromes of multiple endocrine neoplasia)
- Lymphoma

Other neoplasms

- Secondary deposits from extrathyroidal tumours

BOX 19.7 Clinical features suggestive of malignancy in patients with thyroid nodules

- Age <20 years or >80 years
- Hoarseness or voice change
- Breathlessness or stridor from upper airway obstruction
- Rapid enlargement of nodule
- Cervical lymphadenopathy
- Family history of thyroid cancer
- History of radiation exposure
- Smoking history

plasma thyroglobulin is of no diagnostic value, but is essential for long-term monitoring for recurrence.

Treatment

The management of thyroid cancer should be undertaken in a multidisciplinary care setting comprising health professionals with expertise in the field including oncologists, surgeons, endocrinologists, pathologists and radiologists. For papillary and follicular carcinoma, treatment is by total thyroidectomy and, unless the carcinoma is small (<1 cm in diameter), in which case the prognosis is excellent, adjunctive treatment is given with ablative doses of ^{131}I . In order to maximize the effect of the radioiodine, the patient should stop any thyroxine for four weeks beforehand to achieve a serum TSH of >25 mU/L. Alternatively, symptomatic hypothyroidism can be avoided by increasing the serum TSH with the use of recombinant human TSH (rhTSH), thus allowing levothyroxine therapy to continue uninterrupted. Whole body scanning, taking advantage of the therapeutic ^{131}I , will identify any metastatic disease at this stage. Thereafter, thyroxine should be given either in suppressive or replacement doses depending on individual patient risk (Box 19.8). The differentiated thyroid carcinomas are TSH dependent and suppression of TSH has been shown to improve outcomes in patients with high risk disease. Thus, full suppression (TSH <0.1 mU/L) is recommended for patients with persistent disease, while partial suppression (TSH 0.1–0.5 mU/L) is advised in disease-free high-risk patients. TSH suppression is not required in patients with low or very low risk (Box 19.8). Patients on suppressive therapy should undergo periodic cardiovascular and bone mineral density evaluation because of the risk of osteoporosis and atrial fibrillation. Further evaluation should be undertaken six months after initial treatment and should comprise a neck examination, cervical ultrasound and rhTSH-stimulated Tg measurement. Patients without clinical or ultrasound evidence of disease who have normal stimulated Tg concentrations (<1.0 ng/mL) are considered to be in remission and will then require long-term surveillance with annual neck examinations and basal Tg measurements.

Tumour markers

Following total thyroidectomy and ^{131}I ablation, there should be no normal thyroid tissue remaining, and any

BOX 19.8 Target ranges for serum TSH concentration in the follow-up of patients with differentiated thyroid cancer

1. Patients with persistent disease: TSH suppression to <0.1 mU/L
2. Disease-free high-risk patients^a: TSH suppression to 0.1–0.5 mU/L for 5 years
3. Disease-free low-risk patients^a: TSH in low normal range, 0.3–1 mU/L
4. Disease-free very-low-risk patients^a: TSH in normal range, 1–2 mU/L

^aHigh risk: any T3, any T4, any T with N1 or any M1; Low risk: T1–T2, N0, M0 or multifocal T1, N0, M0; Very low risk: unifocal, T1 (<1 cm), N0, M0. T1: tumour <2 cm; T2: tumour 2–4 cm; T3: tumour >4 cm; T4 tumour extending beyond the thyroid capsule or anaplastic histology.

N0: no regional lymph node involvement; N1: regional lymph node involvement; M0: no distant metastasis; M1: distant metastasis.

Modified from Cooper et al. 2009 Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer. *Thyroid* 19:1167–21, and Pacini F et al. 2010 ESMO Guidelines Working Group. Thyroid cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology* 21(Suppl 5):v214–9.

thyroglobulin detected in the serum must arise from the current tumour or from metastases. The sensitivity of the current thyroglobulin assays is poor and the most meaningful results are obtained following TSH stimulation, either as a result of thyroxine withdrawal or by rhTSH stimulation. An undetectable serum thyroglobulin, which fails to increase following TSH stimulation, is indicative of cure and a reason for less intensive follow-up. Response to treatment of recurrent disease or metastases can be monitored by measuring the serum thyroglobulin concentration. The utility of thyroglobulin measurements is limited in the presence of thyroglobulin antibodies, as these antibodies interfere with the thyroglobulin assays. Tests of thyroglobulin recovery have been used in a bid to overcome this problem, but these tests are not entirely reliable and the clinician must be aware of the limitations of thyroglobulin assays in monitoring patients with differentiated thyroid cancer.

Medullary carcinoma of the thyroid secretes calcitonin, which may therefore, be used as a tumour marker. The tumour occurs in both sporadic and familial forms, and because of the association with syndromes of multiple endocrine neoplastic (MEN) disease, patients should be screened for the other manifestations of MEN (e.g. hyperparathyroidism and pheochromocytoma). If the diagnosis of MEN is confirmed, first-degree relatives of the index patient should also be screened for these diseases. Multiple endocrine neoplasia is discussed in detail in Chapter 41.

SYNDROMES OF RESISTANCE TO THYROID HORMONES

These occur most commonly due to a mutation in the nuclear T3 receptor (TR β). The mutation may be inherited (usually autosomal dominant) or occur de novo. Thyroid hormone resistance has been detected in approximately

TABLE 19.8 Overview of factors that enable thyroid hormone resistance to be distinguished from TSH secreting adenomas

Factor	Thyroid hormone resistance	TSH-secreting adenoma
Thyroid status	Usually clinically euthyroid	Usually clinically thyrotoxic
Serum α TSH subunit	Normal	Elevated
Serum SHBG	Often high (85%)	Normal
TSH response to TRH stimulation	Rise in TSH	No response
MRI pituitary	No tumour visualized	Primary tumour visualized (However, incidentaloma detected in 10% of normal subjects)
Genetic analysis of thyroid hormone receptor B gene	Mutation confirmed	No mutation

in 1 in 50 000 live births and affects males and females equally. The condition is characterized by elevated concentrations of total and free thyroid hormones, normal or slightly increased plasma TSH, normal or exaggerated response to TRH and absence of the usual symptoms and metabolic consequences of thyroid hormone excess. Goitre is often present. It is important to distinguish these patients from those with a pituitary TSHoma (Table 19.8). The SHBG concentration is often normal in patients with thyroid hormone resistance, but is increased in TSHoma. Suppression of TSH secretion by high doses of T3 is suggestive of thyroid hormone resistance rather than a TSHoma. A useful pointer to thyroid hormone resistance is to obtain thyroid function test results in first-degree relatives. DNA analysis to demonstrate the mutation in the *TR β* gene should be performed to confirm the diagnosis. Some patients appear to have a relatively selective pituitary resistance and present with hyperthyroid-type symptoms. Management is complex and should be reserved for specialist centres. In many cases no treatment is required.

SCREENING

Is screening for thyroid disease justified? There are undoubtedly some groups of patients in whom thyroid function testing is justified, even if there is no clinical evidence of thyroid disease. The question of screening in neonates has already been discussed. Screening is warranted in all patients with a personal history of previous thyroid disease or neck irradiation, ophthalmic Graves disease, atrial tachyarrhythmias or a history of goitre or previous thyroiditis. In view of the increased prevalence of long-term thyroid dysfunction in patients who have previously received antithyroid drugs, this group of individuals should also be offered regular thyroid function testing. Screening patients with a personal or family

history of autoimmune disorders such as type 1 diabetes, pernicious anaemia, coeliac disease, Addison disease and rheumatological disorders will also yield a significant proportion of patients with undiagnosed thyroid disorders. Thyroid screening is also indicated in patients receiving medications which affect thyroid function such as amiodarone, lithium, tyrosine kinase inhibitors, α -interferon, Campath[®] and antiretroviral agents. Screening patients who have hyperlipidaemia (who have a higher prevalence of hypothyroidism than the normal population) is also justified. The situation is far less clear in the case of screening larger sections of the general population, such as the elderly female population or the hospital in-patient population. In the former example, the cost of such a screening programme has to be balanced against the likely pick-up rate, and in the latter, the confounding influence of non-thyroidal illness is likely to reduce the value of screening in this circumstance. Screening programmes should probably be targeted on otherwise well but 'at-risk' individuals in the community, rather than on the hospital population at large.

The value of thyroid screening in all pregnant women is contentious. Most international endocrine associations do not recommend universal thyroid screening but propose screening only in pregnant patients at risk of thyroid disease, e.g. women with a personal or family history of thyroid disease, co-existent autoimmune disease, goitre, thyroid antibodies, history of head and neck irradiation, or history of poor obstetric outcomes. However, such a targeted case finding approach would miss a third of pregnant women with thyroid disease. A randomized controlled trial of a thyroid screening strategy in early pregnancy showed no beneficial effect on childhood neurointellectual development at three years.

ACKNOWLEDGEMENT

We would like to thank Geoffrey J. Beckett and Anthony D. Toft who wrote the chapter in the previous edition of this book.

Further reading

- Abalovich M, Amino N, Barbour LA et al. Management of thyroid dysfunction during pregnancy and postpartum: an endocrine society clinical practice guideline. *J Clin Endocrinol Metab* 2007;92(Suppl. 8):S1–S47.
- Braverman LE, Utiger RD, editors. *Werner and Ingbar's the thyroid. A fundamental and clinical text.* 9th ed Philadelphia: Lippincott Williams & Wilkins; 2005.
- A comprehensive text covering all aspects of the thyroid.* British Thyroid Association. UK Guidelines for the use of thyroid function tests, http://www.british-thyroid-association.org/info-for-patients/Docs/TFT_guideline_final_version_July_2006.pdf [Accessed October 2013].
- Cooper DS, Biondi B. Subclinical thyroid disease. *Lancet* 2012;379:1142–54.
- Dayan CM. Interpretation of thyroid function tests. *Lancet* 2001;357:619–24.
- Ekins R. Measurement of free hormones in blood. *Endocr Rev* 1990; 11:5–46.
- Ferrara AM, Onigata K, Ercan O et al. Homozygous thyroid hormone receptor beta-gene mutations in resistance to thyroid hormone: three new cases and review of the literature. *J Clin Endocrinol Metab* 2012;97:1328–36.

- Franklyn JA, Boelaert K. Thyrotoxicosis. *Lancet* 2012;379:1155–66.
- Okosieme OE. Thyroid hormone replacement: current status and challenges. *Expert Opin Pharmacother* 2011;12:2315–28.
- Pacini F, Castagna MG, Brillì L et al. ESMO Guidelines Working Group. Thyroid cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2010;21(Suppl. 5):v214–9.
- Shivaraj G, Prakash BD, Sonal V et al. Thyroid function tests: a review. *Eur Rev Med Pharmacol Sci* 2009;13:341–9.
- Warner A, Mittag J. Thyroid hormone and the central control of homeostasis. *J Mol Endocrinol* 2012;49:R29–35.
- Woodmansee WW, Haugen BR. Uses for recombinant human TSH in patients with thyroid cancer and nodular goiter. *Clin Endocrinol (Oxf)* 2004;61:163–73.

Metabolic response to stress

Robin Berry • Philip Gillen

CHAPTER OUTLINE

INTRODUCTION 403

THE RESPONSE TO STRESS 403

Initiation of the stress response 403

Hypothalamo–pituitary–adrenal axis 404

Adrenal medulla 405

STRESS AND THE KIDNEYS 406

CYTOKINES 406

STRESS AND INFLAMMATION 406

Catecholamines 407

Acute phase proteins 408

Coagulation factors 408

SHOCK 409

CARE OF THE SHOCKED PATIENT 409

Definitions 409

Management 410

CONCLUSION 411

INTRODUCTION

Stress can be defined as any influence, arising either internally or externally to the body, that threatens to disrupt normal structure, function or behaviour. Not surprisingly, complex mechanisms have evolved to protect the body against such threats, whether physical or psychological. It may be surmised that during man's evolution, the most important stresses were external, e.g. attack and lack of food, water or shelter. For many people, such factors are now of little or no relevance, but others, for example psychological stresses, have become more prevalent; stressors such as trauma and infection remain universal.

The body's response to stress tends to be qualitatively similar whatever the stress, although the severity of the stress affects the response quantitatively, albeit with considerable intra- and interindividual variation.

It is reasonable to suppose that the development of the metabolic response to stress was subject to evolutionary pressures, and that the response is potentially beneficial. For some aspects this clearly is the case, but, as will become apparent, some aspects of the response to severe stressors, particularly sepsis, appear potentially harmful to the body as a whole. One can only speculate how this might have come about. The metabolic response to stress is complex, and includes processes that act to propagate the response as well as to curtail it and thereby limit its extent, both physiologically and anatomically.

Conventionally, stressors are divided into four categories:

1. physical, e.g. heat, cold, immobilization and pain
2. psychological, e.g. anxiety and fear
3. social, e.g. bereavement, marital breakdown
4. cardiovascular and metabolic, e.g. exercise, haemorrhage and infection.

While this system helps categorize the nature of the stimulus, multiple stressors can be applied to an individual at any one time and these, in turn, can be either acute or chronic in nature. Stress responses are intended to be short lived and self-limiting, but are proportional to the intensity of the stimulus and, consequently, can range from simple localized reactions to complex processes. Systemic responses involve vascular, endocrine and metabolic changes orchestrated by the hypothalamo–pituitary axis and sympathetic nervous systems. Centrally, there is facilitation of arousal, inhibition of vegetative functions (feeding and reproduction) and activation of counter-regulatory feedback loops. Respiratory and heart rates increase and there is an increase in vascular tone, resulting in a rise in blood pressure and hence oxygen and nutrient supply to the brain, heart and skeletal muscles.

The intensity of stress may be gauged by peak concentrations of stress hormones and neurotransmitters, by physiological changes, such as increases in heart rate and blood pressure, and by the length of time for which these changes persist during stress and following the cessation of stress. Whilst intended to confer a survival advantage, the stress response can become part of the pathological process. In severe stress, responses that appear potentially advantageous if limited in extent (e.g. mobilization of muscle protein to provide glucose as an energy source) become harmful.

THE RESPONSE TO STRESS

Initiation of the stress response

Figure 20.1 summarizes the complexity of the systemic responses to potential stimuli. Stimulation of the

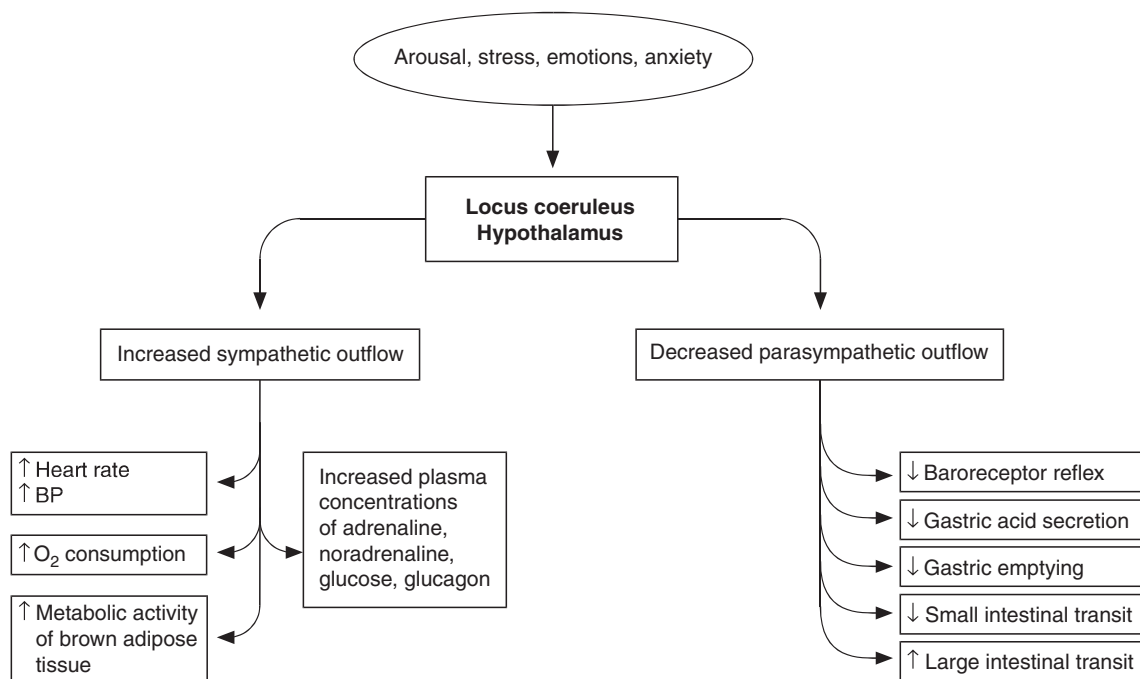


FIGURE 20.1 ■ The systemic response to stress.

hypothalamus and locus coeruleus leads to the release of pituitary hormones and activation of the sympathetic nervous system. The locus coeruleus, hypothalamus and brainstem are closely linked, both anatomically and functionally. The locus coeruleus and hypothalamus mutually innervate and stimulate each other. This positive feedback system allows for activation of the systemic stress response by any stimulus that initiates either side of this loop. The locus coeruleus is located in the lateral floor of the fourth ventricle, and represents the major pool of noradrenaline (norepinephrine) secreting neurons in the brain. It receives afferents from many areas, including the hypothalamus, the cerebellum, prefrontal cortex and the hypoglossi. The cingulate gyrus and amygdala also innervate the locus coeruleus, transmitting emotional and pain stressors. Nerve fibres from this nucleus innervate the spinal cord, brainstem, thalamic relay nuclei, cerebellum and hypothalamus and are predominantly excitatory in action. Adrenaline (epinephrine) secreting nerve terminals are also found within the locus coeruleus and are thought to represent part of the medullary regulatory circuit.

At the centre of the limbic system, the hypothalamus controls most of the vegetative (homeostatic) and endocrine functions of the body. Combined with the locus coeruleus, it is the major effector of the stress response.

The efferent neuronal pathways from the hypothalamus can be classified into two groups: direct monosynaptic projections to pre-ganglionic neurons in the lower brainstem and spinal cord projections to brainstem catecholaminergic neurons in the medulla that innervate neurons in the thoracic spinal cord.

Additionally, hypothalamic nuclei participate in other regulatory pathways, such as temperature control, energy homeostasis and body fluid composition.

Hypothalamo–pituitary–adrenal axis

Stimulation of the paraventricular nucleus by afferent fibres from the limbic system and lower brainstem results in the secretion of corticotrophin-releasing hormone (CRH), into the primary capillary plexus of the hypophyseal portal system, from where it is transported to the anterior pituitary gland.

The hypothalamus also secretes the polypeptide arginine vasopressin (AVP), which is transported along nerve tracts to the posterior pituitary. Under resting conditions, CRH and AVP are secreted in a circadian and highly concordant pulsatile fashion: stressors promote a surge in AVP secretion. Vasopressin increases water retention in the kidneys, is a potent vasoconstrictor, acts as a neurotransmitter and has the ability to modulate a number of physiological processes including gluconeogenesis, platelet aggregation and inflammation.

Corticotrophin-releasing hormone and AVP act synergistically on the anterior pituitary to induce synthesis of the large prohormone, proopiomelanocortin (POMC). The POMC molecule is processed to form a series of smaller peptides, the nature of which is dependent on the enzyme systems present within the tissue type. Anterior pituitary corticotrophin cells express prohormone convertase 1 (PC1), which cleaves POMC into ACTH, N-terminal peptide, joining peptide and β -lipotropin. In the hypothalamus the enzyme PC2 cleaves POMC to α -, β - and γ -melanocyte stimulating hormones and β -endorphin.

α -Melanocyte stimulating hormone acts on melanocortin receptors in the paraventricular nucleus of the hypothalamus, stimulating the sympathetic nervous system. Adrenocorticotrophic hormone and the other peptide cleavage products of POMC have a negative feedback effect on the hypothalamus to limit CRH release.

While CRH and AVP act synergistically on the anterior pituitary, only small amounts of ACTH are secreted in the absence of CRH, which is pivotal to an integrated stress response.

Cortisol

Following stimulation by ACTH, cortisol is the principle glucocorticoid produced by the adrenal cortex, accounting for 95% of glucocorticoid activity. Approximately 90% of circulating cortisol is protein bound; some is attached non-specifically to albumin, but the majority binds with high affinity to cortisol-binding globulin (CBG); only free cortisol is biologically active. Cortisol-binding globulin is an α_2 -globulin, produced in the liver, which behaves as a negative acute phase protein. The concentration of CBG falls in response to some inflammatory stimuli, increasing the bioavailability of cortisol.

Cortisol is a major effector of the metabolic response to stress. It is lipid soluble and readily crosses cell membranes, its activity being mediated via the glucocorticoid receptor (GR). Two isoforms of GR have been isolated, GR α , which binds cortisol and the GR β , which does not bind cortisol and may act to inhibit the effect of endogenous glucocorticoid action. The hormone receptor complex is translocated to the nucleus where it interacts with glucocorticoid response elements and, following recruitment of co-activators, leads to increased gene transcription. Other actions are DNA independent and involve direct interactions with transcription factors, such as nuclear factor- κ B (NF- κ B). This is important in the modulation of the stress response, since NF- κ B activates the expression of interleukin-1 (IL-1), IL-6 and tumour necrosis factor α (TNF α), all of which are key components of the inflammatory response and interact with the hypothalamo-pituitary-adrenal axis (HPAA).

Cortisol exerts a negative feedback effect on CRH and ACTH release, affects carbohydrate, protein and fat metabolism, and modulates the immune response. Cortisol limits glucose utilization in cellular respiration, perhaps by reducing the oxidation of NADH to NAD⁺ in the tricarboxylic acid cycle. High concentrations of cortisol reduce the sensitivity of peripheral tissues to insulin. This anti-insulin effect is particularly marked in skeletal muscle and adipose tissue, reducing glucose uptake and metabolism.

In extra-hepatic tissues, there is increased protein catabolism and decreased protein synthesis through reduced RNA formation and amino acid uptake. Within the liver, however, this metabolic picture is reversed, with increased amino acid uptake and the synthesis of acute phase proteins. Additionally, cortisol enhances the mobilization of fatty acids from adipose tissue, shifting oxidative metabolism to favour energy generation from glucose to fatty acids.

Thyroid hormones

Major disturbances of thyroid hormone metabolism and plasma concentration occur during the metabolic response to stress, typically with falls in the concentrations

of tri-iodothyronine and often of thyroxine and, sometimes, in the critically ill, a fall in the concentration of thyroid stimulating hormone. (These changes are discussed further in Chapter 19.)

Sex hormones

The effects of trauma and sepsis on sex hormones are complex and dependent on the type of injury. Plasma concentrations of free testosterone may fall without any consistent change in the concentration of luteinizing hormone (LH), suggesting a change in the sensitivity of Leydig cells to LH or a disturbance of normal feedback regulation. The effects tend to be more marked in women, with reduced activity of the entire hypothalamo-pituitary-gonadal axis. Corticotrophin releasing hormone is known to have inhibitory effects at various levels in this axis, including antagonizing the effects of LH on Leydig cells. Chronic activation of the hypothalamo-pituitary-gonadal axis, such as has been demonstrated in long distance runners and ballet dancers, causes suppression of gonadal function in males and females.

Growth hormone

The anterior pituitary secretes growth hormone, stimulated by growth hormone releasing hormone, produced by the ventral medial nucleus of the hypothalamus. Growth hormone is a small protein that is active in most cells of the body, where it enhances amino acid uptake and protein synthesis. Under the influence of growth hormone, free fatty acids are used as an energy source in preference to carbohydrates and proteins. Growth hormone reduces glucose uptake by skeletal muscle and fat cells, promotes gluconeogenesis by the liver and consequently increases insulin secretion by the pancreas.

Adrenal medulla

Activation of the locus coeruleus by stressors causes the release of noradrenaline and increased sympathetic discharge via the brainstem and spinal cord. Pre-ganglionic sympathetic nerves pass through the sympathetic chain to innervate modified neuronal cells in the adrenal medulla. These cells secrete noradrenaline and adrenaline into the circulation. Catecholamine synthesis is an energy-dependent process with many mitochondria packing the terminal varicosities containing the secretory vesicles. Tyrosine is converted to L-DOPA and then to dopamine, which is transported into the secretory vesicles where a hydroxylation reaction converts it into noradrenaline. The majority of noradrenaline is then metabolized to adrenaline. The sympathetic nerve impulse to the adrenal medulla results in the influx of intracellular calcium and the discharge of the catecholamine secretory vesicles (in the ratio of 80:20, adrenaline:noradrenaline) directly into the circulation. The combined effect of this circulating adrenaline and noradrenaline is identical to that caused by direct sympathetic nervous system stimulation, but has a longer duration and can reach all cells of the body irrespective of their autonomic innervation.

STRESS AND THE KIDNEYS

Fluid conservation occurs as part of the stress response, often resulting in low urine outputs despite euvoelaemia. Increased sympathetic nervous system activity, combined with raised concentrations of circulating catecholamines, reduces renal blood flow. This initiates the metabolism of prorenin to renin in the juxtaglomerular cells of the kidneys. Renin is released into the blood where it acts on angiotensinogen to form angiotensin I. Under the influence of angiotensin-converting-enzyme, predominantly in the lung, angiotensin I is converted to angiotensin II, a potent vasoconstrictor. Angiotensin II stimulates AVP release from the hypothalamus, as well as having a direct affect on the kidney to increase salt and water retention. Furthermore, under the influence of angiotensin II, the mineralocorticoid aldosterone is synthesized in the zona glomerulosa of the adrenal cortex, by metabolism of corticosterone by the enzyme aldosterone synthase. Whilst ACTH has a permissive role in aldosterone synthesis, it does not control the rate of release. Aldosterone acts on the cortical collecting tubules where it stimulates a Na^+, K^+ -ATPase, leading to the retention of sodium and water. (Salt and water balance is considered further in Chapter 4 and renal function in Chapter 7.)

CYTOKINES

Cytokines are peptides secreted by many cell types including mast cells, macrophages and endothelial cells. They are released into the extracellular fluid and can function as autocrine, paracrine or endocrine hormones and they trigger the acute phase response.

Major changes in the plasma concentrations of cytokines occur during trauma and sepsis: they are also released during psychological stress. Cytokines are produced by activated cells of the immune system and mediate communication between the immune, endocrine and central nervous systems. Cytokines are classified into three groups dependent on their function. Group 1 (IL-2, IL-3, IL-4, IL-7, IL-10, IL-11, IL-12 and granulocyte-macrophage colony stimulating factor) act on a variety of cell types as positive or negative growth factors, whilst group 2 (TNF α / β , IL-1 α / β , IL-6, IFN- α / γ , IL-8 and macrophage inhibitory protein-1) have proinflammatory properties. Group 2 cytokines are responsible for the induction of fever, muscle catabolism, activation of white blood cell precursors and growth of inflammatory fibroblasts and macrophages. Group 3 contains cytokines with anti-inflammatory activity (IL-1 receptor antagonists, soluble IL-1 receptors, TNF α binding protein and IL-1 binding protein). Endothelial cells and macrophages are major producers of IL-1, IL-6 and TNF α , but virtually all nucleated cells have the ability to synthesize them.

Physical damage to cells results in the release of cytokines, some of which are chemotactic (e.g. IL-8), attracting large numbers of macrophages and neutrophils to the site of injury, where they are activated by TNF α , IL-1 and IL-6 to phagocytose dead tissue and bacteria.

Whether cytokines have localized (paracrine) or generalized (endocrine) effects is an important determinant of

the extent of the inflammatory response. Cytokines having local actions are also released into the circulation where they initiate some of the systemic features of the metabolic response, such as fever (IL-1) and stimulation of hepatic protein synthesis to produce acute phase proteins (IL-6): IL-6 is also a potent stimulator of ACTH (and hence cortisol) secretion, which amplifies the stress response.

Interleukin-1, IL-6 and TNF α are particularly important cytokines. Interleukin-1 is produced by activated macrophages and mediates the inflammatory response, stimulates the HPA and potentiates amino acid flux in skeletal muscle. Tumour necrosis factor- α , produced by macrophages and lymphocytes has similar properties to IL-1, and acts synergistically with it. Tumour necrosis factor- α stimulates inducible nitric oxide synthase, causing vasodilatation whilst also stimulating the production of cyclooxygenase, which in turn facilitates prostaglandin and thromboxane synthesis. Prostaglandin E_2 acts on the hypothalamic thermoregulatory centre to cause fever and is also immunosuppressive; other prostaglandins excite A and C pain fibres. Thromboxanes stimulate platelet function and are procoagulant. Interleukin-6 may also contribute to the activation of the HPA and to insulin resistance. Tumour necrosis factor- α , IL-1 β and IFN- γ are crucial for the induction of other cytokines (e.g. IL-6 and IL-8), platelet activating factor, prostaglandins, leukotrienes and nitric oxide.

STRESS AND INFLAMMATION

Stress and glucocorticoid hormones can either enhance or suppress immune function, depending on the following factors:

- the duration of the stress
- changes in leukocyte distribution and the compartments involved
- the concentration, duration and nature of the glucocorticoid exposure
- the timing of stress or stress hormone exposure relative to the stage of an immune response.

The stress response promotes a physiological synchronization of cardiovascular, musculoskeletal and neuroendocrine systems to confer a survival advantage. Despite the known anti-inflammatory effects of glucocorticoids, the body is unlikely to suppress immune function when the host is liable to wounding and infection. Indeed it is more likely that acute stress may serve to enhance the immune response.

Tissue injury (trauma, infection, hypoxia) causes the release of proinflammatory mediators such as heat shock proteins, adenosine and high mobility group box protein 1 in a characteristic damage-associated molecular pattern, and they are known collectively as 'alarmins'. Effective immunoprotection requires rapid recruitment of leukocytes to the site of damage, which is facilitated by the binding of alarmins to endogenous receptors on neutrophils and macrophages. Subsequently, the activation of NK- κ β leads to the transcription of proinflammatory proteins, including the cytokines IL-1, IL-6 and TNF α . Following release, these cytokines further activate inflammatory cells and the vascular system leading to a systemic response.

Cortisol and adrenaline act synergistically to induce rapid changes in leukocyte distribution and to enhance cell mediated and contact hypersensitivity immunity. Stress-induced changes include a reduction in the numbers of lymphocytes and monocytes but an increase in the number of neutrophils. These changes are rapidly reversed upon cessation of the stress.

Glucocorticoids act synergistically with cytokines to enhance specific immune reactions, by inducing their release and potentiating their actions. Synergistic interactions between glucocorticoids and cytokines may be mediated by glucocorticoid induced upregulation of cytokine receptors on target cells as determined by increased cytokine binding or cytokine receptor mRNA expression. For example, glucocorticoid-induced TNF receptors have been shown to promote survival and serve as a stimulatory receptor for T cells. Furthermore, glucocorticoids increase IL-1 binding to human peripheral blood B cells.

Acute psychological stress is known to increase circulating TNF α , IL-1 β and IL-10, with glucocorticoids also enhancing the biological activity of IL-2, interferon γ (IFN- γ) and granulocyte colony stimulating factor. Acute stress responses may induce a biphasic shift in leukocyte numbers. Initially, catecholamine hormones and neurotransmitters induce leukocytes to exit the spleen and bone marrow and to enter blood vessels and lymphatics. Consequently, there is an increase in blood leukocytes, predominantly natural killer (NK) cells and granulocytes. As the stress response evolves, glucocorticoid hormones initiate migration of lymphocytes to sites of infection. This redistribution leads to a reduction in the numbers of circulating lymphocytes, but is not an anti-inflammatory effect.

In contrast to acute stress, chronic stress has been shown to dysregulate immune responses by altering the cytokine balance from type 1 to type 2. Immunosenescence is accelerated and immunity decreased owing to a reduced number of circulating protective cells and increased populations of suppressor T lymphocytes.

To limit potential tissue damage after activation of the immune system, a counter-regulatory response is simultaneously triggered by inflammation and hypoxia. An important physiological role of endogenous glucocorticoids might be to suppress an ongoing immune response. Studies have demonstrated the existence of a negative feedback loop between the immune system and HPA axis such that proinflammatory mediators arising from ongoing immune reactions stimulate the axis, which in turn results in the secretion of cortisol, which suppresses the immune response and prevents it from potentially damaging the host.

Cortisol exerts its anti-inflammatory effects through a number of mechanisms. By stabilizing lysosomal membranes, cortisol reduces the release of proinflammatory proteolytic enzymes and reduces damage to the vascular endothelium, thereby decreasing capillary leak. Cortisol is thought to have an anti-inflammatory action through the non-genomic binding of the GR-cortisol complex to the transcriptase NK-k β , reducing the formation of prostaglandins and cytokines that would act as positive chemotactic agents for phagocytic white blood cells. Cortisol

inhibits the production of IL-2, which stimulates the proliferation of T cells. It also induces the synthesis of lipocortin, an inhibitor of phospholipase A2, which is required for the synthesis of prostaglandins and leukotrienes. Concurrently, there is secretion of IL-6 an anti-inflammatory cytokine that inhibits the production of IL-1 and TNF α and thus provides negative feedback for the inflammatory cascade.

Cortisol can promote increased transcription of anti-inflammatory genes. Mitogen activated protein kinases (MAPKs) form intracellular signalling pathways that result in the production and post-transcriptional modification of proinflammatory mediators. MAPK phosphatases dephosphorylate MAPKs rendering them inactive; glucocorticoids increase the expression of MAPK phosphatases and thus can down regulate the inflammatory response.

Cells that surround the inflamed tissues express specific tissue protective receptors (TPR) that bind the type 1 cytokine erythropoietin. Binding of this ligand inhibits pro-inflammatory cytokine production, inhibits macrophage activity and delimits the volume of injury by countering apoptosis. Activation of TPRs also acts to recruit vascular and tissue specific stem cells and enhances tissue repair.

Catecholamines

Noradrenaline is released mainly from the postganglionic nerve endings, while the adrenal medulla is the major source of adrenaline in the circulation.

Stimulation of the sympathetic nerves to the adrenal medulla results in large quantities of adrenaline and noradrenaline being released into the circulation. The action of these catecholamines is mediated via cell surface adrenergic receptors, which are themselves bound to protein molecules that traverse the cell membrane. Binding to these receptors causes a conformational change in the protein, resulting in the opening of an ion channel or the stimulation of a second messenger enzyme system. There are two major groups of adrenergic receptors, α and β , which are further divided into subgroups (α_{1-2} , β_{1-3}). Noradrenaline preferentially excites α receptors in addition to having some β activity, whilst adrenaline is equally effective across both receptor groups. Sympathetic stimulation is excitatory in some tissues and inhibitory in others and similarly α and β receptors may be excitatory or inhibitory.

Circulating catecholamines enhance the sympathetic nervous system's stimulation of the cardiovascular system. This results in arteriolar and venous constriction and increased heart rate and force of myocardial contraction. The net result is raised ventricular blood volume causing increased muscle fibre stretching, improving cardiac output and raising blood pressure.

During stress, noradrenaline and adrenaline both cause hyperpolarization of gut smooth muscle and decreased bowel motility. Circulating adrenaline stimulates the phosphorylase enzyme via cAMP, causing rapid glycogenolysis especially in the liver and skeletal muscle. Combined with the anti-insulin effects of cortisol, adrenaline and noradrenaline activate triglyceride lipase in fat cells, causing the rapid mobilization of fatty acids, shifting energy metabolism away from glucose.

Brain catecholaminergic systems utilize noradrenaline, adrenaline, dopamine and L-DOPA. Stressful stimuli accelerate both the synthesis and release of brain catecholamines, with adrenergic and noradrenergic neurons involved in the central processing of the stress response. The effect of these neurotransmitters is dependent on which receptors and transporters are expressed on target cells. Adrenaline binds to the same receptors as noradrenaline and adrenergic neurons may use noradrenaline transporters for reuptake from synaptic clefts.

Acute stress leads to transient alterations in brain catecholamine systems with transient activation of the HPA. After a short period of time and depending on the nature of the stressor, the activity of these catecholaminergic neurons can return to baseline. Chronic, long-term or repeated stressors may induce permanent changes and keep catecholamine producing neurons chronically active.

In response to the majority of stressors, brainstem noradrenaline producing neurons stimulate the HPA non-specifically via adrenergic receptors. In contrast, stimulation of the paraventricular nucleus of the hypothalamus is stressor specific.

Acute phase proteins

Hepatic protein synthesis shifts from constitutive proteins (e.g. albumin and transferrin), to acute phase proteins (e.g. C-reactive protein (CRP) and fibrinogen), following activation of the systemic stress response. This causes increased plasma concentrations of CRP, coagulation proteins, protease inhibitors and complement proteins. These acute phase proteins are arbitrarily divided into type I, e.g. haptoglobin and α_2 -acid glycoprotein that are mediated by IL-1 and TNF-like cytokines, and type II acute phase proteins mediated by IL-6-like cytokines. Since the synthesis of this group of proteins increases under stress, they are referred to as positive acute phase proteins. In contrast, constitutive proteins such as albumin and the hormone binding proteins are downregulated and classified as negative acute phase proteins. Under stress conditions, the amino acids required for this protein synthesis are derived from extrahepatic tissues, where proteolysis and reduced amino acid uptake occur, secondary to the influence of cortisol and the catecholamines.

Positive acute phase proteins have diverse functions in opsonization and trapping of microorganisms, activating complement, binding cellular remnants, neutralizing enzymes, scavenging free radicals and in modulating the host's immune response.

C-reactive protein is a ring-shaped protein that binds to phosphocholine on the surface of dead cells and some bacteria where it acts as an opsonin, activating complement by the classical pathway. Haptoglobin binds free haemoglobin, inhibiting its oxidative activity and has anti-inflammatory capabilities, binding to major receptors on the cell membranes of leukocytes. It is removed from the circulation, along with any bound molecules, by the reticuloendothelial system. Fibrinogen (factor 1) is a high molecular weight glycoprotein involved in the clotting cascade and inflammation. It is converted by the enzyme thrombin into fibrin monomers that automatically

polymerize to form fibrin fibres as part of clot formation. Fibrinogen and fibrin contribute to inflammation by inducing leukocyte migration and by modulating the inflammatory response of leukocytes and endothelial cells via an increased responsiveness to cytokines and chemokines.

Coagulation factors

Coagulation may be enhanced during the stress response to infection, with the inflammatory and clotting cascades being closely linked. Endothelial cell injury activates the clotting cascade through binding of factor VIIa to tissue factor, leading to the generation of thrombin. This catalyses the formation of fibrin and activates platelets, resulting in the formation of a clot. During stress, there are increased concentrations of circulating procoagulants such as fibrinogen and the cytokine IL-6, which stimulate the expression of tissue factor. Concurrently, platelets are activated by proinflammatory cytokines and adrenaline. There is downregulation of physiological anticoagulant mechanisms such as proteins C and S (negative acute phase proteins) and inhibition of fibrinolysis. Consequently, inflammation-induced coagulation is characterized by widespread intravascular fibrin deposition. Impairment of thrombolysis occurs because of high circulating concentrations of plasminogen activator inhibitor type 1.

By binding to protease activated receptors (PARs), thrombin induces leukocyte migration to the surface of the endothelium and by induction of the expression of cytokines and chemokines by leukocytes. The four isoforms of PARs (PAR 1–4) belong to a family of G-coupled receptors activated by proteolytic cleavage.

Procoagulant activity is regulated by three important anticoagulant pathways: antithrombin, the protein C system and tissue factor pathway inhibitor. During inflammation-induced coagulation, the function of all three pathways can be impaired.

The serine protease inhibitor antithrombin is the main inhibitor of thrombin. During severe inflammatory responses, antithrombin concentrations are markedly decreased owing to impaired synthesis (negative acute phase response), degradation by elastase from activated neutrophils and consumption by ongoing thrombin generation.

Antithrombin has an anti-inflammatory effect, by stimulating prostacyclin release from endothelium. The anti-inflammatory effects of cytokines are the result of blockade of both cytokine production by neutrophils and their tethering to vascular endothelium. Antithrombin also binds to specific receptors on neutrophils, monocytes and lymphocytes, blocking their interactions with the endothelium.

In severe inflammation, plasma concentrations of protein C are low due to impaired synthesis, consumption and degradation by proteolytic enzymes such as neutrophil elastase. There is also a significant downregulation of thrombomodulin, a protein cofactor expressed on the surface of endothelial cells. Under normal conditions, the thrombin–thrombomodulin complex activates protein C, but in severe inflammation the expression of this protein is inhibited by the inflammatory cytokines TNF α and IL-6. Low plasma concentrations of free protein S may compromise the function of the protein C system.

In plasma, 60% of the cofactor protein S is bound to complement regulatory protein C4b binding protein (C4bBP). Increased plasma concentrations of C4bBP, as a consequence of the acute phase response, result in a relative protein S deficiency. In sepsis, the endothelial protein C receptor is downregulated, further impeding the function of this anticoagulant.

SHOCK

Shock can be classified according to its aetiology: cardiac dysfunction (either primary or secondary); loss of circulating volume (fluid loss or redistribution), or failure to maintain systemic vascular resistance. These may occur in combination, particularly when entering the final common pathway of refractory shock. If untreated, the end result, whatever the cause, is tissue hypoperfusion and impairment of cellular metabolism.

The amount of oxygen delivered to tissues is a function of the cardiac output and arterial oxygen content. In haemorrhagic shock, cardiac output falls because of reductions in preload, force of contraction and stroke volume. These changes promote compensatory sympathetic responses inducing tachycardia, peripheral vasoconstriction, tachypnoea and the activation of neurohumoral mechanisms. The adrenal glands secrete adrenaline, noradrenaline and cortisol, whilst renal hypoperfusion activates the renin–angiotensin system. Angiotensin II, a vasoconstrictor, stimulates the release of vasopressin and aldosterone. Vasopressin, a pressor agent, increases the permeability of the renal collecting ducts to water, resulting in water retention, whilst aldosterone acts at the distal renal tubule to promote sodium retention.

If the stress response is inadequate for the metabolic challenge imposed upon the body, endothelial cells switch to anaerobic metabolism, absorb interstitial fluid and swell. This causes narrowing of the lumen of arterioles, further restricting the oxygen supply. Ischaemic cells accumulate lactic acid and free radical species, which are directly cytotoxic, prompting the release of inflammatory mediators and the adhesion of neutrophil leukocytes cells. As the inflammatory cascade is propagated, the cumulative anaerobic metabolites become negatively inotropic, and energy-dependent potassium channels fail, causing arterioles to dilate, heralding the onset of refractory shock.

CARE OF THE SHOCKED PATIENT

The response to stress is typically well regulated but homeostatic control can be lost in the presence of powerful stimuli or if a second insult occurs, for example an infection or a need for surgery after trauma.

Definitions

Patients are admitted to critical care units because of various underlying disorders, yet may display similar clinical manifestations regardless of the clinical insult. This observation led to the introduction of the term 'systemic inflammatory response syndrome' (SIRS) and

BOX 20.1 Definitions used in critically ill patients

Infection: a microbial phenomenon characterised by an inflammatory response to the presence of micro-organisms or the invasion of normally sterile host tissues or fluids by these organisms

Systemic inflammatory response syndrome (SIRS): two or more of the following:

- temperature $>38^{\circ}\text{C}$ or $<36^{\circ}\text{C}$
- heart rate >90 beats/minute
- respiratory rate >20 breaths/minute or $\text{PaCO}_2 <4.3\text{kPa}$
- white cell count $>12 \times 10^9$ or $<4 \times 10^9$ or $>10\%$ immature forms

Severe sepsis: infections causing SIRS with associated organ failure, hypoperfusion or hypotension (systolic BP <90 mmHg)

Septic shock: severe sepsis with arterial hypotension refractory to fluid replacement

to attempts to define shock in order to establish a consistent basis to assist diagnosis and treatment (Box 20.1). Such a classification also facilitates design of clinical trials and incorporation of their results into routine clinical practice. The term multi-organ dysfunction syndrome (MODS) is now preferred to the previously used term 'multi-organ failure'. Multi-organ dysfunction syndrome is defined by organ specific criteria, involving clinical and investigation findings. It is now preferred as a concept as it is acknowledged that it may be appropriate to offer organ support before the function of individual organs has deteriorated to a life-threatening extent. Signs of organ system dysfunction are shown in Box 20.2.

BOX 20.2 Signs of organ system dysfunction

Neurological dysfunction: altered mental status (reduced Glasgow Coma Scale)

Respiratory dysfunction: arterial hypoxaemia ($\text{PaO}_2 <10\text{kPa}$) and/or hypercapnia ($\text{PaCO}_2 >6.5\text{kPa}$); the presence of infiltrates radiologically and the requirement for respiratory support

Cardiovascular dysfunction: ischaemia, tachydysrhythmias, systolic hypotension (<90 mmHg) and a requirement for vasopressor and/or inotropic support. Increased capillary refill times and hyperlactataemia (>2 mmol/L)

Renal dysfunction: acute oliguria (urine output <0.5 mL/kg/hr) with increased serum creatinine concentration, acid base and electrolyte disturbances

Hepatic dysfunction: elevated liver enzymes, hyperbilirubinaemia and coagulopathy; hypoglycaemia in severe cases

Gastrointestinal dysfunction: intolerance of enteral feeding, ileus, gastrointestinal bleeding and ischaemia

Haematological dysfunction: leukocytosis or leukopenia, thrombocytopenia and platelet dysfunction, disseminated intravascular coagulation

Endocrine dysfunction: hypothalamo–pituitary–adrenal axis failure; hyperglycaemia, hypertriglyceridaemia

Immunological dysfunction: hospital-acquired infection

Multi Organ Dysfunction Syndrome (MODS): when two or more organs are involved

Management

The mortality and morbidity of critically ill patients has been much improved by the Surviving Sepsis Campaign (SSC), an international forum launched in 2004 to improve the management of sepsis. Since the majority of patients with SIRS, shock and MODS have infection as a contributory factor, appropriate management of sepsis is essential.

Immediate care

Immediate measures include removing the source of infection (if practicable), collection of tissues and fluids for microbiology and the commencement of antimicrobials. Targets in this phase of management include a heart rate <90 bpm, systolic BP >90 mmHg (MAP >65 mmHg), a urine output >0.5 ml/kg/min and normalisation of serum lactate concentration (if previously raised).

Initial antimicrobial therapy must be empirical, on the basis of the likely causative organisms until the results of culture and sensitivity are known. Factors to consider include the site of the infection, knowledge of the patient's pre-morbid condition, previous antibiotic exposure and local prescribing guidelines. In the case of antibiotics, broad-spectrum agents can often be replaced with a tailored regimen once sensitivities are known. This approach reduces the risk of promoting multi-drug resistance in pathogenic bacteria. Gram staining of specimens may reveal bacteria and recently the application of PCR (polymerase chain reaction) technology has led to rapid diagnostics in some infections.

Organ support

Balanced salt solutions (e.g. Hartmann's, lactated Ringer's) are the fluids of choice in the resuscitation of septic shock. Norepinephrine (noradrenaline) is typically the favoured vasopressor (because of its predominant α -adrenergic effects), often in combination with the vasodilating inotrope, dobutamine. In cases of refractory shock, vasopressin is a valuable and potent pressor agent, usually used in combination with noradrenaline and intravenous steroids. Frequently, the use of vasopressin will allow a reduction in the dose of noradrenaline, potentially limiting some of its deleterious effects. Glucocorticoids enhance vascular responsiveness to catecholamines by mechanisms that may include adrenoceptor function and prostaglandin action. Hydrocortisone is prescribed at a dose of 200 mg daily in divided aliquots or as a continuous infusion. Whilst it is not necessary to perform a tetracosactide test to identify adrenal suppression, hydrocortisone should not be commenced in septic patients who are not shocked, because of its immunosuppressive effects.

Respiratory support should be given as required. Patients receiving mechanical ventilation should undergo intermittent interruptions to assess the need for ongoing assistance. If their underlying condition permits, they should also be nursed with the head of the bed elevated to 30–45° as this has been shown to limit aspiration and reduce the risk of pneumonia.

Red cell transfusion should generally be reserved for patients whose haemoglobin concentration falls to <70 g/L, once tissue hypoperfusion has resolved, except in particular circumstances such as acute haemorrhage or myocardial ischaemia. Platelet transfusion should be administered prophylactically in severe sepsis when platelet counts are <10 × 10⁹/L, in the absence of severe bleeding and 50 × 10⁹/L in the context of active bleeding, surgery or invasive procedures. Intravenous immunoglobulins are not recommended in severe sepsis or septic shock.

Continuous renal replacement therapy and intermittent haemodialysis have been shown to be of equal benefit in patients with severe sepsis. However, continuous therapies facilitate fluid balance in patients who are haemodynamically unstable.

In the first 48 h after a diagnosis of severe sepsis or septic shock, it is preferable to administer nutrition enterally, as tolerated, rather than to fast the patient or to provide only intravenous glucose. Current evidence suggests that intravenous glucose and low dose enteral feeding, increased as tolerated, is preferable to parenteral feeding in the first week of illness. Various immunomodulating agents have been used in nutritional formulations but none is currently recommended for use in patients with sepsis. The reduction in the availability of arginine in sepsis has the potential to lower nitric oxide synthesis and increase production of superoxide and peroxynitrite. Conversely, arginine supplementation has the potential to cause vasodilatation and hypotension and studies of its use in nutritional preparations have had conflicting results. Plasma glutamine concentration is reduced during severe illness. A previous meta-analysis showed a reduction in mortality associated with its use but this has not been reflected in other studies. Whilst there do not appear to be adverse effects of glutamine supplementation, a large trial (the REDOXS study) is currently in progress to investigate its use. There is some initial evidence that the administration of selenium may have a beneficial effect in SIRS and shock by providing an antioxidant defence. Further studies are required before its administration can be recommended for this specific purpose, but its use as part of standard trace element supplementation is endorsed.

There should be protocols in place for the management of blood glucose. Initially, blood glucose should be monitored every 1–2 h and insulin infusion begun if two consecutive concentrations are >10 mmol/L. Once glucose concentrations and insulin infusion rates are stable, the frequency of monitoring may be reduced to every four hours. The upper acceptable limit for blood glucose should be 10 mmol/L in this group of patient; the stricter target of 6.1 mmol/L advocated previously was not associated with reduced mortality, but predisposed to hypoglycaemia.

Patients with severe sepsis should receive mechanical prophylaxis against venous thromboembolism using graduated compression stockings or mechanical compression devices. In addition, they should receive pharmacoprophylaxis with low molecular weight heparin unless contraindicated, e.g. thrombocytopenia, active bleeding, recent intracerebral haemorrhage.

Patients with severe sepsis and risk factors for bleeding (e.g. coagulopathy, mechanical ventilation for at least 48 h and possibly hypotension) are at risk of stress ulceration and should be given prophylaxis using proton pump inhibitors in preference to H₂ receptor blockers.

Immunomodulation

The body has a non-specific, or innate, immune response provided by mechanical barriers such as the skin and gastric acid secretion, and a more specific acquired immune response. Both may be altered by treatment, either directly or as a secondary effect, e.g. proton pump inhibitors given for stress ulceration reduce gastric acidity (potentially compromising innate immunity) and nutritional support may have a beneficial effect on the immune response.

As the pathogenesis of SIRS, shock and MODS involves immune mediators, it is logical to attempt to treat the conditions by their specific modulation. Various agents have been tried for this purpose.

The use of glucocorticoids for their vasopressor effect in shock has been described above. Their effect on the immune response is via a reduction in the production of proinflammatory cytokines, particularly TNF α and IL-6, although trials of the use of glucocorticoids for immunomodulation have not shown great benefit. Similarly, trials of anti-cytokine therapies, such as monoclonal antibodies to TNF β and recombinant IL-1 receptor antagonist, have not proven their efficacy.

Inflammatory and procoagulant responses are intimately linked and may lead to widespread injury to vascular endothelium, contributing to MODS. Protein C acts to promote fibrinolysis and has inhibitory actions on thrombosis and inflammatory responses which make it a potential therapeutic target. Activated protein C has been used in the treatment of severe sepsis but has now been withdrawn due to lack of evidence of benefit on 28-day mortality.

Some agents used routinely in the critically ill, e.g. sedatives such as propofol and benzodiazepines, and non-steroidal anti-inflammatory agents, have been shown, particularly in animal studies, to modify the immune response, although a precise clinical role is unclear. Recent evidence suggests that statins may modify the immune response but the results of clinical trials to evaluate this further are awaited.

One explanation for the lack of clinical benefit from immunomodulating therapies may be the redundancy of action that exists between the cytokines; blocking of one pathway still allows for activation and release of mediators by an overlapping route. A further reason could be that the timing of the administration of the immunomodulating agent may be of critical importance for the success of a response.

CONCLUSION

The body's response to stress tends to be qualitatively similar whatever the stress, although the severity of the stress affects the response quantitatively. In severe stress, responses that appear potentially advantageous if limited in extent become harmful. There have been improvements in the survival of critically ill patients as a result of better organ support and more effective treatment of infection, but attempts at specific immunomodulation of the inflammatory response have been disappointing.

Further reading

- Dellinger RP, Levy MM, Rhodes A et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. *Crit Care Med* 2012;41:580–637.
- Kvetnansky R, Esther L, Sabban EL et al. Catecholaminergic systems in stress: structural and molecular genetic approaches. *Physiol Rev* 2009;89:535–606.

Disorders of puberty and sex development

S. Faisal Ahmed • Jane D. McNeilly

CHAPTER OUTLINE

INTRODUCTION 412

NORMAL SEX DEVELOPMENT 412

NORMAL PUBERTAL DEVELOPMENT 414

Endocrinology of normal puberty 414

Physical signs of normal puberty 414

DISORDERS OF SEX DEVELOPMENT 416

Terminology of disorders of sex development 416

General principles of management 417

General examination of a newborn with suspected DSD 417

Evaluation of the external genitalia 417

Evaluation of the internal anatomy 418

Investigating the newborn with DSD 418

Investigating the adolescent with DSD 419

Steroid measurement and its interpretation 420

Anti-Müllerian hormone 420

Insulin-like factor 3 420

Inhibins 421

The human chorionic gonadotrophin (hCG) stimulation test 422

The role of the clinical geneticist 422

Classification of disorders of sex development 423

DISORDERS OF PUBERTY 428

Precocious puberty 428

Gonadotrophin dependent puberty (central causes) 428

Delayed puberty 429

Hypogonadotrophic hypogonadism 430

Primary hypogonadism 431

INTRODUCTION

The clinician is most commonly faced with the investigation and management of abnormal sexual development at two age periods: birth and puberty. Abnormal sexual development at birth is the result of a disorder of sex development (DSD) presenting with genitalia that are either atypical for sex or ambiguous. Disorders of sex development can also present at adolescence as abnormalities of pubertal development. However, in most cases, abnormal pubertal development is due to an alteration in the regulation of puberty in an individual without a DSD. To understand these disorders there is a need to understand the normal processes of sex and pubertal development.

NORMAL SEX DEVELOPMENT

Male or female sex development is programmed at fertilization, depending on whether the zygote is heterogametic (46XY) or homogametic (46XX). A scheme summarizing the embryology and genetic control of sex

development in the male and female fetus is shown in [Figure 21.1](#). The indifferent, or primitive, gonad develops from a thickened ridge of coelomic epithelium that arises from the intermediate mesoderm. Cell precursors from this region give rise to the kidneys, adrenals and gonads. The adrenogenital primordium comprises a single population of steroidogenic factor-1 (SF-1) immunoreactive cells. The gonadal ridge becomes populated with primordial germ cells that have migrated from the wall of the yolk sac.

The first histological expression of fetal testicular development is the appearance of seminiferous or sex cords at between six and seven weeks of gestation. Surrounding the sex cords is the interstitial region, which contains the precursors of Leydig cells and also the peritubular myoid cells. Primitive interstitial cells differentiate later into Leydig cells, an event that coincides with the onset of testosterone synthesis by the fetal testis. Steroidogenesis is placental human chorionic gonadotrophin (hCG)-dependent and results in the production of fetal testosterone concentrations within the normal adult male range. The pivotal role for the Y chromosome in male sex differentiation has been recognized for more than

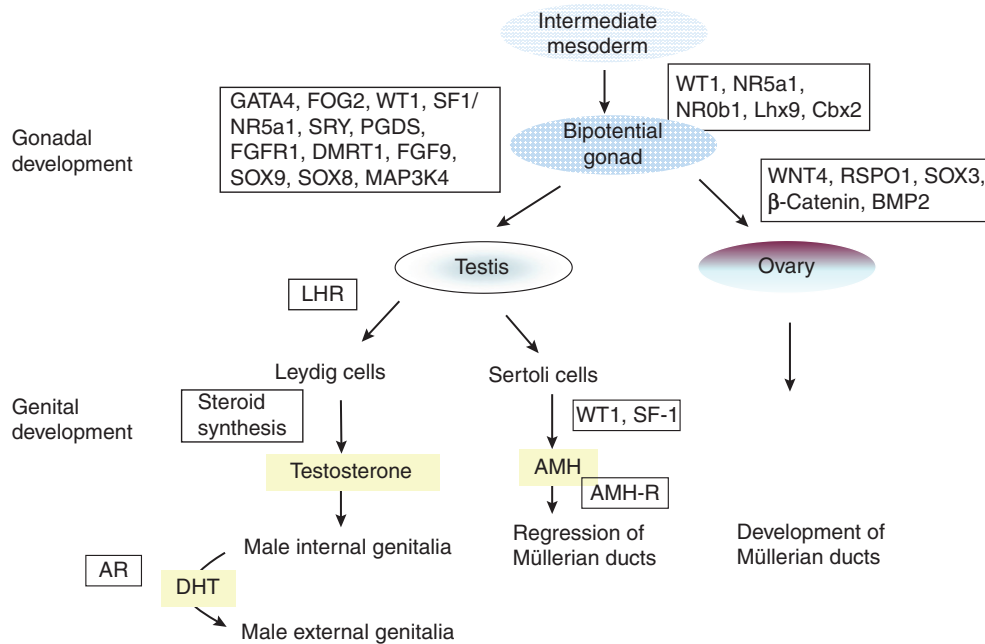


FIGURE 21.1 ■ Embryology and genetic control of fetal sex development. LHR, LH receptor; AR, androgen receptor; DHT, dihydrotestosterone; AMH-R, anti-Müllerian hormone-receptor.

40 years, based on the phenotype of individuals with sex chromosome aneuploidies such as XXY and XO. The study of sex-reversed XX males and XY females led to the identification of the *SRY* gene (sex-determining region Y) on the short arm of the Y chromosome. That this is the prime testis-determining gene is illustrated by the introduction of the *Sry* homologue into a transgenic XX mouse and a resulting male phenotype. The majority of XX males are positive for the *SRY* gene, as a result of terminal exchange between the X and Y chromosomes during paternal meiosis. In the case of XY females, mutations have been identified that disrupt the function of the *SRY* protein, which acts as a transcription factor. However, these are found in only 15–20% of patients, indicating there are other genes also involved in testis determination. *SRY* acts to up-regulate the transcription of the related gene, *Sox9* in the Sertoli cell lineage. High concentrations of *SOX9* protein are maintained in these cells through a positive interaction with the secreted molecules fibroblast growth factor 9 (FGF9) and prostaglandin D2 (PGD2). These molecules are important in securing two other necessary steps in testis development: the recruitment of other somatic cells to differentiate into the Sertoli cell lineage and, in the case of FGF9, spread of the initial central testis-determining signal to the anterior and posterior poles of the gonad. Not only does *SOX9* promote testis development, but it also antagonizes the Wnt signalling pathway and *FOXL2*, both of which are crucial for ovarian development and which can, themselves, antagonize *SOX9* action.

With the committed development of the gonad as a testis, two trophic factors produced by the testis control subsequent differentiation of internal and external genitalia in the male. Anti-Müllerian hormone (AMH) is produced by the Sertoli cells and signals through two trans-membrane receptors expressed in the mesenchyme that gives rise to

the Müllerian ducts. Anti-Müllerian hormone acts ipsilaterally to cause regression of the Müllerian ducts, the anlage (rudimentary precursor) of the uterus, fallopian tubes and upper two-thirds of the vagina. Circulating concentrations of AMH in males are high in early infancy and fall gradually throughout childhood, until they become undetectable after puberty; its measurement is useful for detecting testicular tissue, especially in early childhood. Testosterone is the other key trophic factor produced by the early developing testis. There is indirect evidence, based on luteinizing hormone (LH) receptor mouse knockout models and human LH β -chain mutations, that production is initially autonomous, but, thereafter, is under the control of hCG and then fetal pituitary gonadotrophin secretion. High local concentrations of testosterone stabilize Wolffian duct development in the male to differentiate into the vas deferens, epididymis and seminal vesicles. The external genitalia in both sexes develop from a common anlage, comprising the urogenital sinus, genital tubercle and swellings and urethral folds. In the male, androgens, again, play the key role in differentiation of these primordial structures, to develop as the penis, scrotum and opening of the urethra on the glans. Dihydrotestosterone (DHT) is necessary to provide an amplification of the androgenic effect. The evidence for DHT dependency is illustrated by the predominantly female external genital phenotype in 5 α -reductase deficiency.

In the absence of the testis-determining factors, testosterone and AMH, and in the presence of the trophic proteins necessary for ovarian development, the gonad associated with the 46XX karyotype develops as an ovary and the Müllerian ducts differentiate normally into female internal genital ducts. The external genital anlage, in the absence of androgens, remains underdeveloped with respect to growth of the genital tubercle and the absence of midline fusion of the labioscrotal folds and swellings.

The development of the lower part of the vagina from canalization of the vaginal plate is incompletely understood, but the process is probably influenced by oestrogens. Virilization of the female external genitalia from an extraneous source of androgens leaves the development of the female internal genitalia intact.

NORMAL PUBERTAL DEVELOPMENT

Endocrinology of normal puberty

The precise trigger that initiates the onset of puberty in the human remains an enigma. There is activation of pituitary gonadotrophin secretion both in early fetal life and, again, after birth for several months. During this latter period, plasma concentrations of gonadotrophins and sex steroids may reach values normally observed at puberty. Throughout later infancy and childhood, gonadotrophin concentrations remain low, although evidence of pulsatile LH secretion using more sensitive immunoassays can be detected in some children. Consequently, puberty represents the reactivation of gonadotrophin secretion that has been restrained during childhood.

The initial endocrine event at puberty is an increase in nocturnal pulsatile LH secretion in response to gonadotrophin releasing hormone (GnRH) released into the pituitary portal system. The neurons that release GnRH migrate to the hypothalamus, from the medial olfactory placode, during fetal development. A failure of such migration is considered to be the cause of some inherited forms of hypogonadotrophic hypogonadism that are associated with anosmia (Kallman syndrome). A GnRH pulse generator controls the onset of puberty by causing a progressive increase in amplitude and frequency of LH pulses, which occurs about one year before the onset of physical signs of puberty. The neuroendocrine mechanisms that cause activation of the GnRH generator at the appropriate age are extremely complex, but appear to involve excitatory amino acids, catecholamines, neuropeptide Y, leptin and acetylcholine, as well as external influences such as nutritional intake. Furthermore, an additional factor discovered as a result of studies in patients with hypogonadotrophic hypogonadism, is the kisspeptin/GPR54 system, which appears to be a gatekeeper of reproductive development. GPR54 is a G-protein-coupled receptor whose endogenous ligand is a 54-amino acid peptide, cleaved from the parent 145-amino acid peptide encoded by the *KISS-1* gene. GPR54-deficient mice are infertile and have low circulating gonadotrophin concentrations. Mutations in the *GPR54* gene cause hypogonadotrophic hypogonadism in humans. It appears that GPR54 is one of the key factors in the control of puberty. Administration of kisspeptin stimulates LH and follicle stimulating hormone (FSH) secretion, which can be blocked by GnRH antagonists, indicating that the kisspeptin/GPR54 system effect on the hypothalamic-pituitary-gonadal axis is mediated via GnRH release.

Increased gonadotrophin secretion leads to a gradual increase in the secretion of gonadal steroids, principally testosterone in males and oestradiol in females. Random, daytime sex steroid measurements are seldom predictive

of pubertal events in an individual, although early morning testosterone concentration as a reflection of nocturnal secretion is more useful. Inhibin B, a marker of Sertoli cell function, increases early in puberty and reaches adult concentrations by mid-puberty. Spermatogenesis starts between 11 and 15 years of age and sperm can be detected in early morning urine specimens by 13 years of age.

The attainment of reproductive capacity in the female is dependent on cyclical gonadotrophin secretion to ensure ovulation. Follicle stimulating hormone-induced oestrogen secretion, by developing ovarian follicles, leads to a positive feedback on LH secretion and a mid-cycle surge that induces rupture of the mature follicle. After ovulation, the luteinized follicle secretes progesterone while oestradiol concentrations fall. Inhibin B secretion predominates during the follicular phase, while inhibin A secretion is dominant during the luteal phase. All these events are coordinated by pulsatile release of GnRH secretion and modulation of gonadotrophin and sex steroid production, by both negative and positive feedback loops. It is not surprising that there is generally a 1–2-year interval after menarche before the majority of girls are ovulating regularly. Ultrasonography can be used to assess the development of reproductive function in girls. Enlargement of the uterus and endometrial thickening are evident after birth because of the effect of maternally derived oestrogens. Multiple ovarian follicular cysts are often seen on ultrasound at this time. The effect of increasing secretion of oestradiol during puberty can be observed by appropriate morphological changes in the appearance of the uterus. The prepubertal uterus starts to increase in size from seven years of age onwards.

The onset of increased production of dehydroepiandrosterone (DHEA) and its sulphate (DHEAS) by the zona reticularis of the adrenals, between six and eight years of age, defines the phenomenon of adrenarche. Concentrations of DHEAS thereafter increase throughout puberty into early adult life in micromolar amounts and then start to decline gradually in older age (adrenopause). There is no concomitant increase in adrenal glucocorticoid and mineralocorticoid secretion, so ACTH is not the primary trophic factor controlling DHEA production. Intra-adrenal modulation of steroidogenesis, by post-translational regulation of the 17-hydroxylase enzyme (by phosphorylation on serine/threonine residues, and electron transfer by the P450-oxidoreductase enzyme), is the key factor in the production of C19 steroids such as DHEA. There is also a relative underexpression of the enzyme 3 β -hydroxysteroid dehydrogenase enzyme in the zona reticularis that contributes to a preponderance of Δ^5 steroids. Extra-adrenal factors postulated to play a role in adrenarche include prolactin, oestrogens, growth factors and cytokines. There is an association between leptin concentration and the timing of adrenarche in obese children.

Physical signs of normal puberty

Puberty is the transitional period between childhood and adulthood that spans adolescence and leads to the acquisition of reproductive capacity. The time span for

the physical changes to take place is generally 4–5 years, but individual variations can result in a time interval of 2–6 years. This is referred to as the tempo of puberty. The age of onset of puberty also varies considerably, with epidemiological evidence suggesting that puberty may now be starting earlier.

The first sign of puberty in girls is breast development, starting, on average, at 11 years with an age range of 8–13 years (see Fig. 21.2A). This is termed ‘thelarche’ and starts as a small mound of tissue beneath the nipple manifest as a breast ‘bud’, which is usually distinguishable, on palpation, by its firmness compared with the softer and more diffuse texture of subcutaneous fat. Further development of the nipple, areola and underlying breast tissue takes place over the ensuing four years

or so (Tanner stages B2–B5). Epidemiological studies conducted in the USA indicate that 25% of girls may have already started thelarche by eight years of age. However, it is unclear whether misinterpretation of adipose tissue as breast development may have skewed the data.

Coincident with breast development is the onset of growth of pubic hair which may precede thelarche in 10% of girls. Hair growth usually starts on the labia before spreading over the mons pubis and is quantified as Tanner stages PH2–5. Sometimes there may be further growth on the upper medial aspects of the thighs and along the linea alba (PH6). Axillary hair starts to appear at about 12.5 years of age and takes a further 18 months to reach adult distribution.

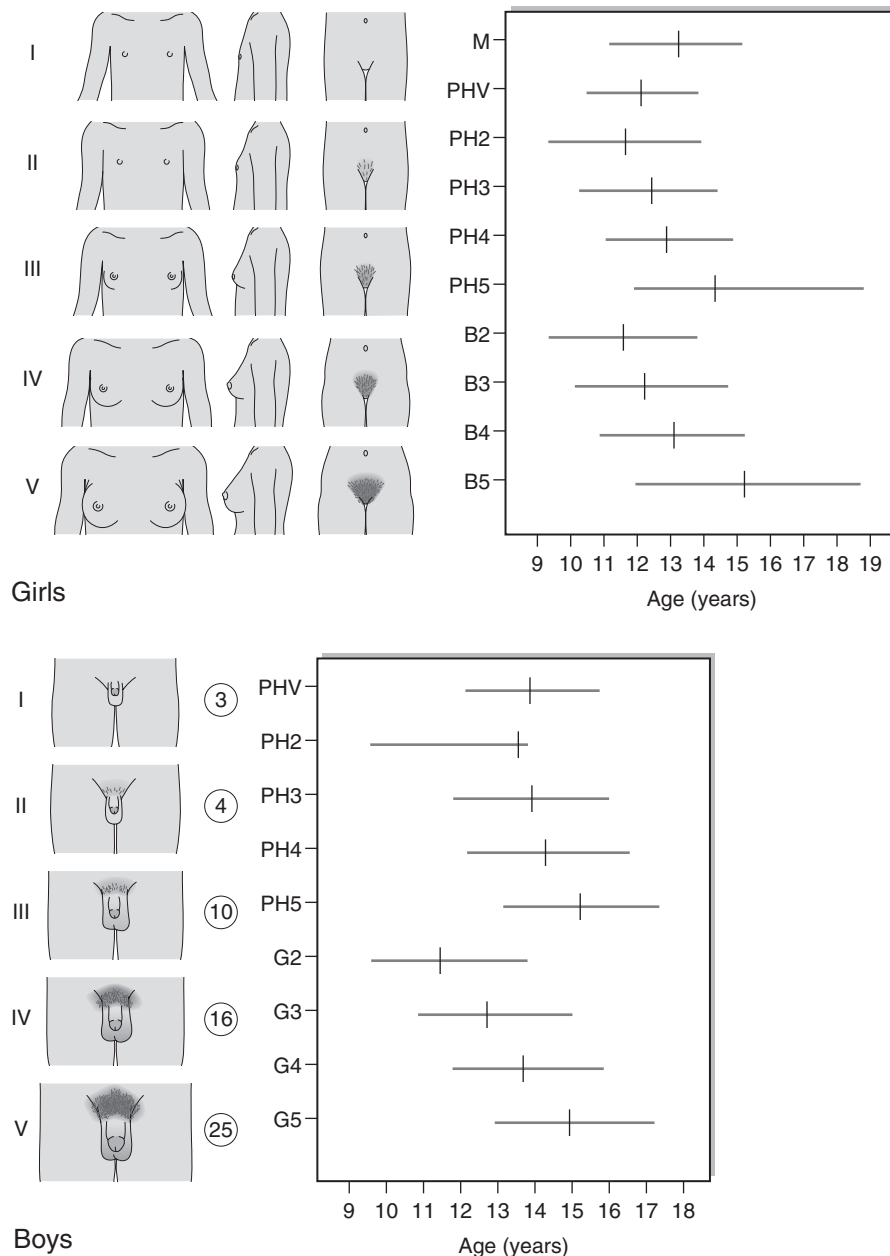


FIGURE 21.2 ■ Tanner staging of physical development, based on external primary and secondary sexual characteristics in children, adolescence and adults. M, menarche; Circled figures, testicular volume (mL); PHV, peak height velocity; PH, B, G, see text.

Menarche, the onset of menses, is a relatively late event in the pubertal process and occurs at around 12.5–13 years of age. This coincides with breast stage B4. A trend in the lowering of the age of menarche has occurred in the past, but the reduction has only been by a few months during the past 30 years. There are racial differences as well as geographical variations. The age of menarche tends to be slightly higher in Northern than Southern Europe.

Increased growth velocity is a measurable, indirect sign of puberty. In girls, peak height velocity (defined as the maximum rate of growth achieved during puberty, which is usually ~10 cm/year) is achieved relatively early in puberty (during breast stages B2–3) and before the onset of menarche. By the time of menarche, the majority of adult height has been achieved. Changes in body composition also occur, particularly with respect to fat distribution.

The first sign of puberty in boys is an increase in testis size (see Fig. 21.2B). Testicular size can be assessed clinically by the use of a series of standard ovoids (the Prader orchidometer). Testicular volume remains between 2–3 mL throughout infancy and childhood, an increase to 4 mL heralds the onset of puberty that occurs, on average, at 11.5 years with an age range of 9–14 years. Progressive enlargement of the testes takes place over three years with testicular volumes reaching up to 25 mL in adult life. Leydig cells constitute only a small part of testicular volume, the majority being the result of an increase in Sertoli cell and seminiferous tubule number and size.

Pubic hair growth, which may start as a few scrotal hairs, follows closely on testicular development and is quantified as Tanner stages PH2–5. Spread of hair up the abdominal wall is characteristic in many men and is classified as stage PH6. Growth of the penis, first in length and then in breadth, occurs concomitantly and is rated G1–G5. Growth of axillary and facial hair, are both later events in puberty, ranging in age from 14 to 16 years. ‘Breaking’ of the voice is due to enlargement of the larynx and elongation of the vocal cords and does not occur until stages G3–G4.

Peak height velocity in boys corresponds with stages G4 and PH4 and testicular volumes of 12–15 mL. This occurs, on average, at a chronological age of 14 years, as opposed to 12 years in girls. The adult male is, on average, taller than the adult female by 13 cm. This is partly due to the fact that boys start the growth spurt later and are, therefore, taller at its onset. In addition, the magnitude of growth achieved during the growth spurt is greater in boys (28 cm in males and 20 cm in females). Body composition alters in favour of increased muscle bulk and relatively less subcutaneous fat.

DISORDERS OF SEX DEVELOPMENT

Terminology of disorders of sex development

It is important to define terms used in relation to the investigation and management of disorders of sex

development that present at birth or in childhood. *Sex assignment* (often used interchangeably with *gender assignment*), is the sexing of an infant at birth as being male or female. This is straightforward in the vast majority; indeed, sex assignment is increasingly an activity that occurs before birth. *Gender identity* refers to how individuals perceive themselves as being male or female, while *gender role* describes characteristics that are sexually dimorphic within a normal population. In childhood, for example, male-typical behaviour is attributed to toy preferences being vehicles and soldiers as opposed to playing with dolls, generally preferred by girls. This rather simplistic illustration of sexual dimorphism has some physiological underpinning from observations of girls’ play behaviour following exposure to androgens in utero. Thus, gender role is more male-typical in girls with congenital adrenal hyperplasia. There is less convincing evidence that prenatal hormones affect gender identity. *Sexual orientation* refers to the subject of erotic arousal, which may be heterosexual, homosexual or bisexual. *Gender dysphoria* or *gender identity disorder* are terms that describe gender dissatisfaction. The phenomenon appears to be more common in individuals with disorders of sex development than in the general population, but it is difficult to identify predisposing factors related to the specific disorder.

The use of terminology that is clear and easy to use and understand by all health professionals, patients and their families is fundamental to the understanding, investigation and management of affected newborns and children. In addition, terminology should respect the individual and avoid terms that might cause offence. The term ‘intersex’ has had variable connotations even among professionals; some employed it as a term that covered all affected newborns while at the other end of the spectrum, some believed that the term should only apply to those where there is complete mismatch between chromosomal and anatomic sex. The consensus reached in Chicago in 2005 on the management of these patients, stressed the importance of terminology and recommended a substitution for the term ‘intersex’ for ‘disorder of sex development (DSD)’, which is defined as any congenital condition in which development of chromosomal, gonadal or anatomic sex is atypical. It also recommended the abandonment of terms such as ‘pseudohermaphroditism’ and ‘true hermaphroditism’. While this new nomenclature (Table 21.1) is easier to use and understand, it will nevertheless evolve over time as our understanding of long-term outcome and molecular aetiology improves. Given that genital anomalies may occur as commonly as 1 in 300 births and may not always be associated with a functional abnormality, some have advocated the use of ‘differences’ in preference to the term ‘disorder’. The strength of the abbreviation ‘DSD’ is that it can be used to cover both differences and disorders of sex development. However, the likelihood of this difference existing as a disorder will depend on the functional implications of the condition, which may be heavily influenced by the social and cultural framework within which the child exists.

TABLE 21.1 Old and new terminologies for abnormal sex development

<i>Previous</i>	<i>Proposed</i>
Intersex	Disorders of sex development (DSD)
Male pseudohermaphrodite. Undervirilization of an XY male, undermasculinization of an XY male	46XY DSD
Female pseudohermaphrodite. Overvirilization of an XX female, masculinization of an XX female	46XX DSD
True hermaphrodite	Ovotesticular DSD
XX male or XX sex reversal	46XX testicular DSD
XY sex reversal	46XY complete gonadal dysgenesis

General principles of management

The management of a child with an abnormality of genital development can often be difficult, particularly in patients for whom the sex of rearing is uncertain. The initial contact with the parents of a child with a DSD is important as first impressions from these encounters often persist. A key point to emphasize is that a child with a DSD has the potential to become a well-adjusted, functional member of society. Establishing a dialogue and building rapport with the affected child and the parents, evaluating the child and then developing a logical, as well as pragmatic, plan for investigations are central to the initial approach and ongoing management. It is paramount that any child or adolescent with a suspected DSD is assessed by an expert with adequate knowledge about the range of variation in the physical appearance of genitalia. If there is any doubt, the patient should be discussed with the regional team. For most patients, particularly in the case of the newborn, the paediatric endocrinologist within the regional DSD team acts as the first point of contact. The underlying pathophysiology of DSD and the strengths and weaknesses of the tests that can be performed, should be discussed with the parents and child, as appropriate, and tests undertaken in a timely fashion. With babies in whom there is true genital ambiguity, it should be explained to the parents that the best course of action may not initially be clear, but the healthcare team will work with the family to reach the best possible set of decisions in the circumstances. Finally, in the field of rare conditions, it is imperative that the clinician shares the experience with others through national and international clinical and research collaboration.

General examination of a newborn with suspected DSD

The physical examination should determine whether there are any dysmorphic features and the general health of the baby. Affected infants, particularly those who have XY DSD, are more likely to be small for gestational age and may display other developmental abnormalities. In

addition, the affected infant should be examined for mid-line defects which may point towards an abnormality of the hypothalamo-pituitary axis. The state of hydration and blood pressure should be assessed, as adrenal steroid biosynthetic defects can be associated with a variable extent of salt loss, masculinization and hypertension. In congenital adrenal hyperplasia (CAH), cardiovascular collapse with salt loss and hyperkalaemia does not usually occur until the second week of life (with salt loss usually evident from day four) and so will not be apparent at birth in the well neonate, but should be anticipated if the diagnosis is suspected. Jaundice (both conjugated and unconjugated) may be observed in babies with hypopituitarism or cortisol deficiency. Urine should be checked for protein as a screen for any associated renal anomaly (e.g. Denys-Drash or Frasier syndromes) and a pre-feed blood glucose concentration should be measured to exclude hypoglycaemia (suggestive of hypopituitarism, or occasionally CAH, e.g. 3β -hydroxysteroid dehydrogenase deficiency). Renal tract anomalies, such as ureteropelvic junction obstruction, vesicoureteric reflux, pelvic or horse-shoe kidney, crossed renal ectopia and renal agenesis, are reported to be more common in children with DSD.

Evaluation of the external genitalia

If the appearance of the external genitalia is sufficiently ambiguous to render sex assignment impossible, or the phenotype is not consistent with prenatal genetic tests, then investigations are clearly required. However, the ability to evaluate external genitalia fully may depend on the expertise of the observer and, before presentation to a specialist, the label of ambiguous genitalia has often already been assigned to newborns where the most appropriate sex of rearing was not clear to those present at the child's birth. The birth prevalence of genital anomalies may be as high as 1 in 300 births but the birth prevalence of complex anomalies that may lead to true genital ambiguity on expert examination may be as low as 1 in 5000 births.

Apart from those whose genitalia are truly ambiguous, infants can often be divided into those who overall seem to have largely male or female genitalia but with some unusual features. However, it is very important to bear in mind that a 46XX newborn infant with congenital adrenal hyperplasia can either present as a girl with clitoromegaly or as a boy with bilateral undescended testes. When evaluating these infants, the clinical features of the external genitalia that require examination include the presence of gonads in the labioscrotal folds, the fusion of the labioscrotal folds, the size of the phallus and the site of the urinary meatus on the phallus, although the real site of the urinary meatus may, sometimes, only become clear on surgical exploration. These external features can be individually scored to provide an aggregate score, the external masculinization score (EMS) (see Fig. 21.3A), or they can be graded according to their overall description as classically described by Prader staging (see Fig. 21.3B).

Infants with suspected DSD who require further clinical evaluation and need to be considered for

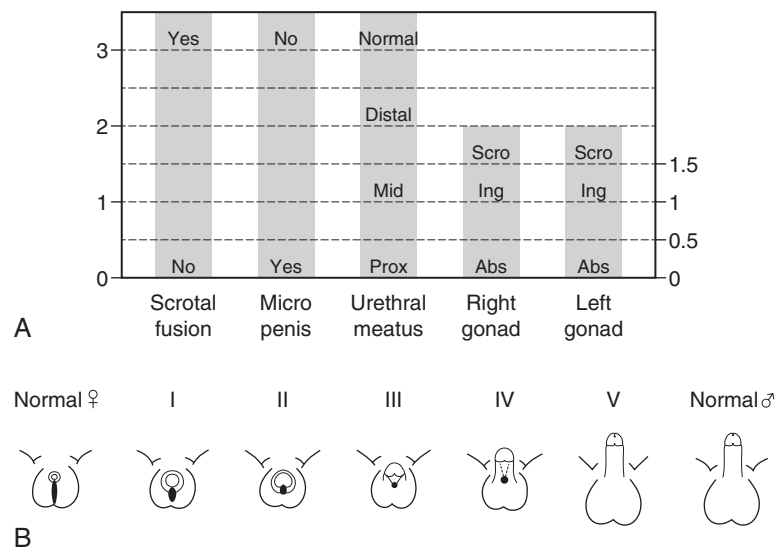


FIGURE 21.3 ■ (A) External Masculinization Score. Each individual feature of the genitalia (phallus size, labioscrotal fusion, site of the gonads and location of urethral meatus) can be individually scored to provide a score out of 12. Microphallus refers to a phallus smaller than normal for age. Scro, scrotal; Ing, inguinal; Abs, abdominal or absent on examination. (B) Differential virilization of the external genitalia using the staging system of Prader from normal female (left) to normal male (right).

investigation by a specialist, should include those with isolated perineal hypospadias, isolated micropenis, isolated clitoromegaly, any form of familial hypospadias and those who have a combination of genital anomalies with an EMS of less than 11. This will avoid unnecessary detailed investigations of boys with isolated glandular or mid-shaft hypospadias and boys with unilateral inguinal testis. The coexistence of a systemic metabolic disorder, associated malformations or dysmorphic features would lower the threshold for investigation as would a family history of consanguinity, stillbirths, multiple miscarriages, fertility problems, genital abnormalities, hernias, delayed puberty, genital surgery, unexplained deaths and the need for steroid replacement. In addition, maternal health and drug exposure during pregnancy, and the pregnancy history itself, may hold key information.

Evaluation of the internal anatomy

Examination and assessment by a paediatric surgeon with experience of DSD is critically important for defining the internal and the external anatomy in the affected patient. Combining examination with endoscopic visualization and radiological assessment can provide information on the location and state of the gonads, the urogenital sinus and Müllerian structures. Ultrasonography is the first-line imaging modality but the reliability is child and operator dependent. In the neonate the uterus, ovaries and adrenals should be identifiable. In the adolescent, it is sometimes difficult to confirm the presence of a prepubertal uterus by ultrasonography and there may be a place for repeat imaging after a six-month course of oestrogen. Magnetic resonance imaging (MRI) may be reserved for patients where ultrasonography has failed to delineate the relationship of the Müllerian structures and where there are abnormalities of the urinary tract. In adolescents, MRI can delineate structural anomalies

such as hydrometrocolpos or hydronephrosis and identify secretory tumours. Nowadays, the 'genitogram' has been superseded by endoscopic examination of the genital tract (genitoscopy), which provides a more detailed and thorough assessment. Laparoscopy is a very effective method of visualizing the internal sex organs and facilitates direct inspection, biopsy or excision of intra-abdominal gonads. In 46XY DSD, laparoscopy is clearly indicated in all children with impalpable testes as the gonads need to be identified and brought down to the scrotum if possible.

Investigating the newborn with DSD

In all infants with ambiguous genitalia and/or bilateral impalpable gonads, a first tier of investigations should be undertaken to define the sex chromosomes, delineate the internal genitalia by pelvic ultrasound and to exclude life-threatening congenital adrenal hyperplasia (CAH) – the commonest cause of ambiguous genitalia of the newborn. This first tier should, therefore, also include plasma glucose, 17OH-progesterone (17OHP), and measurement of sodium and potassium. 17OH-Progesterone measurement is usually unreliable before the age of 36 h and in the salt losing form of CAH, electrolytes usually do not become abnormal before day four of life. The results of FISH (fluorescence in situ hybridization) analysis using Y- and X-specific probes should be available within one working day and the 17OHP results should be available within a maximum of two working days in all specialist DSD centres. In situations where the level of suspicion of CAH is very high and the infant needs immediate steroid replacement therapy, samples should be collected and stored before starting therapy. These should be of a sufficient volume to assess 17OHP, testosterone, androstenedione and, possibly, renin, in that order of priority. At least one spot or 24-h urine sample for a urine steroid profile should be collected before starting therapy. The

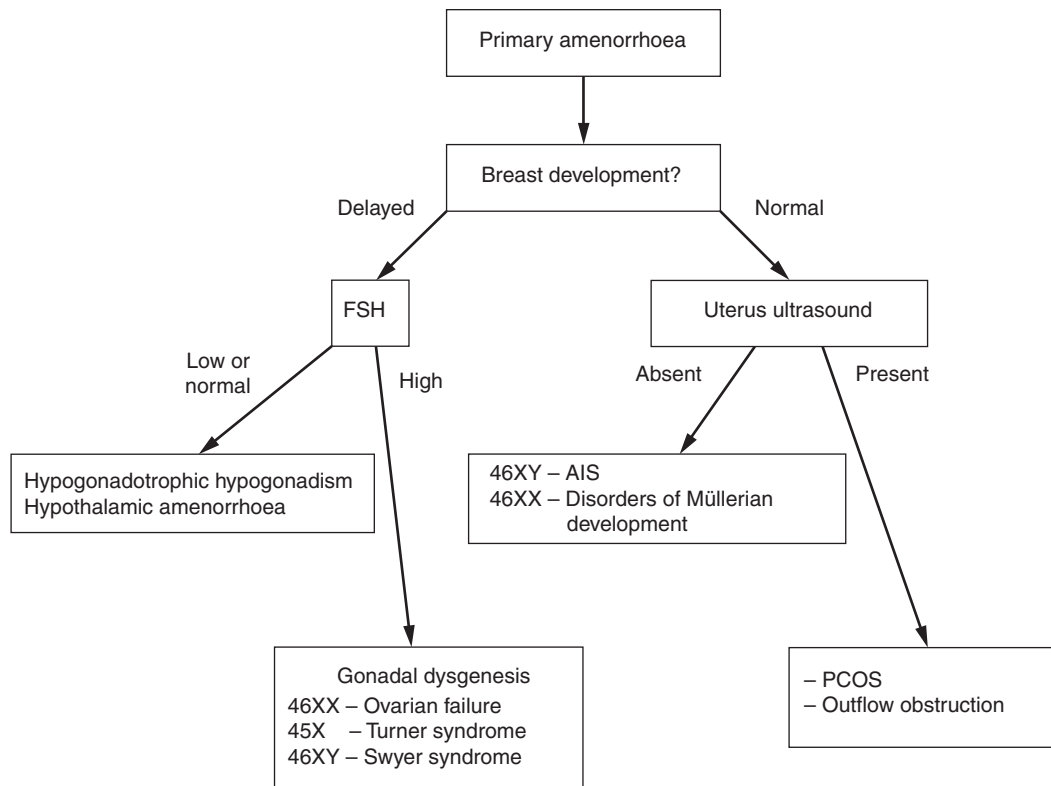


FIGURE 21.4 ■ Approach to investigating adolescent girls with primary amenorrhoea. AIS, androgen insensitivity syndrome; PCOS, polycystic ovary syndrome.

results of these initial investigations will often dictate the second tier of investigations.

In an infant with impalpable gonads, a karyotype of 46XX, a significantly elevated plasma 17OHP concentration and the presence of a uterus, CAH due to 21-hydroxylase deficiency is the most likely diagnosis. A urine steroid profile can confirm this diagnosis and identify other rarer forms of CAH, which may also be associated with a raised 17OHP concentration in the newborn. In infants with sex chromosomes other than 46XX, a second tier of investigations is necessary to determine the presence of testes and the adequacy of androgen production and action. These tests include measurement of AMH, an hCG stimulation test, further imaging and laparoscopy. Confirmation of a specific diagnosis will often require biochemical identification of a defect in the androgen biosynthesis pathway and detailed genetic analysis.

Investigating the adolescent with DSD

Adolescents typically present with a suspected DSD in three ways – as a girl with primary amenorrhoea (with or without breast development), as a girl who virilizes at puberty, or, as a boy with pubertal delay (see Fig. 21.4). The potential psychological impact of a thorough physical examination and medical photography on an adolescent should be carefully considered and, in some circumstances may be appropriate only under an anaesthetic. In girls with primary amenorrhoea, investigations should be considered at the age of 14 years if there is no pubertal development and at 16 years if

other aspects of puberty, particularly breast development, have progressed normally. History should include a family history and an assessment of coexisting chronic disease, exercise and weight changes. Physical examination should include measurement of blood pressure, height and weight and assessment of secondary sexual characteristics including clitoral enlargement. Vaginal examination to assess vaginal length is rarely indicated if imaging is informative and, if carried out, should be clearly explained and performed by a gynaecologist. An initial investigation screen should comprise measurements of LH, FSH, prolactin, thyroid stimulating hormone (TSH), free thyroxine (FT4), sex hormone binding globulin (SHBG), androstenedione, oestradiol and testosterone, and transabdominal pelvic ultrasound, performed by a sonographer with experience of adolescent appearances. Raised gonadotrophins, or an absent uterus in the presence of normal breast development, are indications for karyotyping.

The appearance of clitoromegaly and hirsutism at puberty, in the presence of primary amenorrhoea, is a classical presentation of two 46XY DSDs: 17 β -hydroxysteroid dehydrogenase type 3 deficiency and 5 α -reductase type 2 deficiency. It is less typical of partial androgen insensitivity syndrome (PAIS), which is usually associated with ambiguous genitalia at birth. Müllerian structures will not be detectable in any of these conditions. In partial gonadal dysgenesis and ovotesticular DSD, mild clitoromegaly that may have been present at birth but may have been overlooked, becomes a more prominent feature at adolescence. The differential diagnosis would also

include CAH and androgen-secreting tumours of the ovary or adrenal gland, but in these conditions Müllerian structures are present. Investigations include measurement of LH, FSH, DHEAS, SHBG, androstenedione, testosterone, dihydrotestosterone (DHT) and 17OHP. A 24h urine collection for urinary steroid profiling will confirm 5 α -reductase type 2 deficiency, CAH or adrenocortical tumour. A pelvic ultrasound will assess the presence of a uterus and determine the need for a karyotype.

Although the commonest cause of delayed puberty is constitutional delay, all boys with delayed puberty who are over the age of 14 years should be assessed. Overweight boys need careful examination so that a buried penis is not mistaken for micropenis. Rarely, PAIS, a disorder of testosterone biosynthesis or mild forms of testicular dysgenesis can present in this age group; often there is a history of hypospadias repair or orchidopexy. Investigations include a bone age and measurements of LH, FSH, testosterone and prolactin. For those with raised gonadotrophins, karyotype should be performed to exclude disorders such as Klinefelter syndrome (47XXY and variants) or 45XO/46XY mosaicism.

Steroid measurement and its interpretation

Steroid hormone analysis is a vital component of the biochemical evaluation of the child with DSD but the method of analysis can have a significant impact on the result. Analysis is most often performed by non-extraction (direct) immunoassays on automated platforms and these are subject to concerns of analytical specificity and variability between manufacturers. Liquid chromatography linked with tandem mass spectrometry (LC-MS/MS) allows multiple analyte analysis from a single sample whilst maintaining analytical specificity. Thus, in patients with DSD steroid measurement by either LC-MS/MS or immunoassay after organic solvent extraction is preferred. As these methods tend to be labour intensive, close communication between the clinical and laboratory personnel within the DSD team is vital to ensure timely availability of results.

Urinary steroid profile (USP) analysis by gas chromatography mass spectrometry (GC-MS) provides qualitative and quantitative data on excretion of steroid metabolites (see Fig. 21.5). It is ideal for detecting altered steroid metabolites, especially in patients with CAH, where the activity of a combination of steroidogenic enzymes can produce unusual metabolites which can cross-react in traditional direct serum assays. The diagnosis of rarer forms of CAH such as P450 oxidoreductase deficiency (ORD) is best established using urinary GC-MS analysis as it allows for concurrent determination of all adrenal-derived steroid metabolites. As concentrations of gonadotrophins, androgens and precursors fluctuate markedly over the first few months of life it may be appropriate to consider an early neonatal collection as well as further samples at a later stage. A urine sample can be frozen and stored for many years and may help with a review of the diagnosis at a later stage. Urinary steroid profiling is not appropriate for suspected cases of 5 α -reductase type 2 deficiency until after three months of age as diagnostic ratios of 5 β to 5 α reduced metabolites are not detectable

before then. Infants, particularly boys, normally have significant changes in steroid and other endocrine hormone concentrations during the first 100 days of life. In boys, serum testosterone and DHT may initially be high at birth but decline to <1 nmol/L and undetectable, respectively by around day 30. Concentrations then rise to peak at day 70 before declining to normal prepubertal values. These normal variations may influence the interpretation of sex steroid and gonadotrophin measurements as well as the results of an hCG stimulation test (see p. 422). Furthermore, the actual value for the hormone concentration will vary depending on the assay methodology; it is therefore essential that age and method-related reference ranges are available to facilitate appropriate interpretation of results.

Anti-Müllerian hormone

Anti-Müllerian hormone, also known as Müllerian inhibiting substance, is strongly expressed in Sertoli cells from the time of testicular differentiation to puberty and to a much lesser degree in granulosa cells from birth to menopause. Measurement of AMH concentration in the adult female is used to assess ovarian reserve. Published information on circulating AMH concentrations have to be interpreted with caution owing to differences in primary antibody, standard calibration and units used for measurement in individual immunoassays. In boys, AMH is detectable at birth at much higher circulating concentrations than in girls and these concentrations rise during infancy before gradually declining at puberty. Therefore, it may be helpful to repeat AMH measurement later in infancy if a newborn boy is found to have a low concentration. As summarized in Table 21.2, measurement of AMH is a powerful tool to assess Sertoli cell activity in children with suspected DSD and may also have diagnostic utility in conditions associated with androgen deficiency or insensitivity.

Insulin-like factor 3

Insulin-like factor 3 (INSL3) is one of the hormones secreted by the Leydig cells in the testis, and plays a pivotal role in testicular descent during fetal development. In males, circulating INSL3 concentrations display a characteristic pattern with a transient perinatal increase, low concentrations during childhood, increasing during puberty with highest concentrations found in adulthood when it is expressed constitutively. Insulin-like factor 3 is increasingly being recognized as a more sensitive marker of current Leydig cell differentiation and function than testosterone and, thus, has potential clinical value when used as an adjunct to standard investigations for DSD and may avoid reliance on hCG stimulation tests to assess testicular function adequately. In females, INSL3 is produced by theca cells and the corpus luteum, therefore concentrations are undetectable until puberty, when active ovarian cycling occurs. Circulating INSL3 concentrations are significantly lower in females than males. Lack of age-appropriate reference values, robust, sensitive commercial immunoassays and larger-scale studies are major barriers to the use of INSL3 in routine practice.

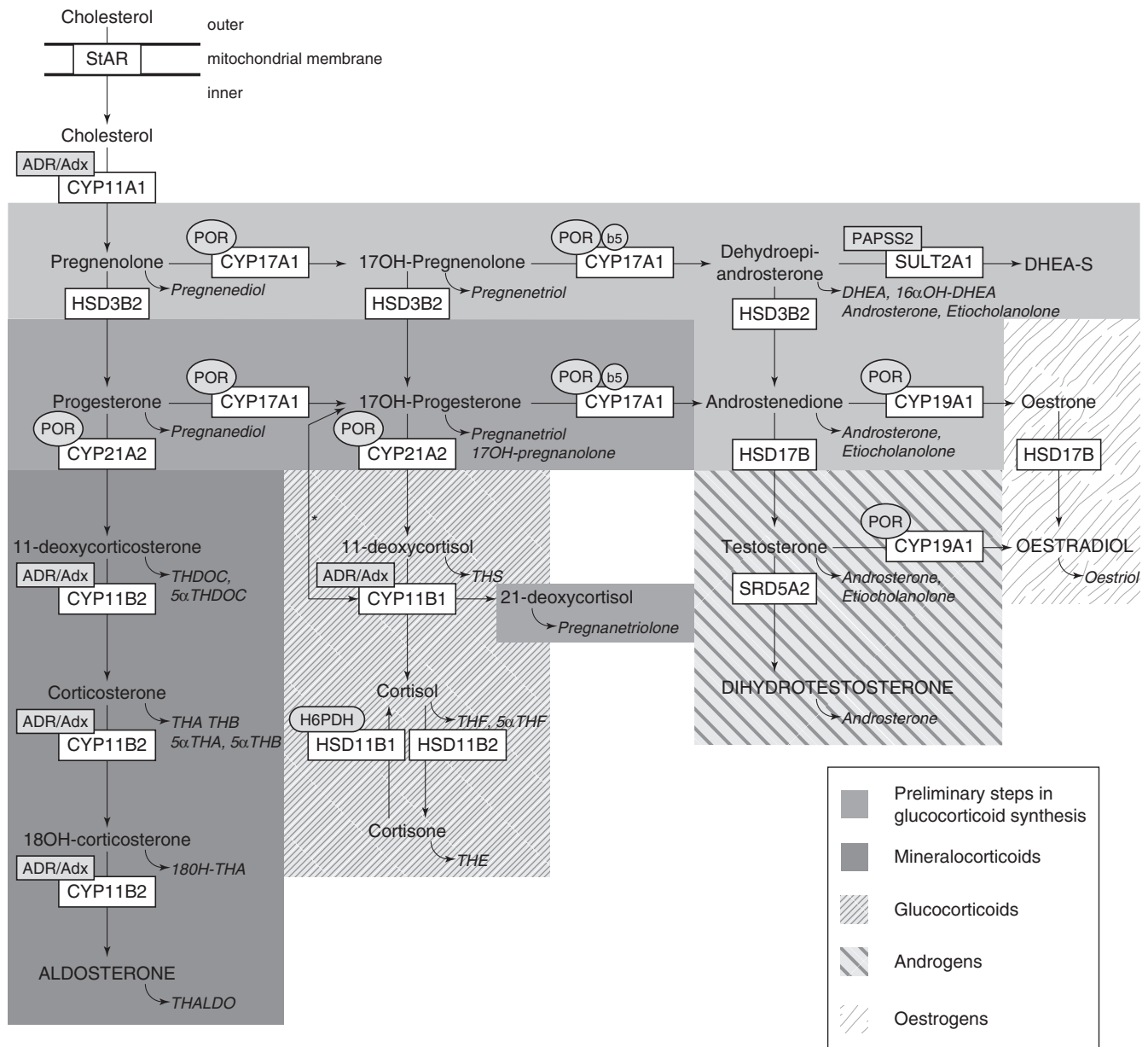


FIGURE 21.5 ■ Synthesis and metabolism of hormonal steroids. This figure illustrates the formation of the major hormone classes from cholesterol. Steroid names in conventional script are steroid hormones and precursors; those in *italics* are urinary metabolites of the aforementioned. The major transformative enzymes are in rectangular boxes, the cofactor ('facilitator') enzymes in ovals. Mitochondrial CYP type I enzymes requiring electron transfer via adrenodoxin reductase (ADR) and adrenodoxin (Adx) CYP11A1, CYP11B1, CYP11B2, are marked with a labelled box ADR/Adx. Microsomal CYP type II enzymes receive electrons from P450 oxidoreductase (POR), CYP17A1, CYP21A2, CYP19A1, are marked by circled POR. The 17,20-lyase reaction catalyzed by CYP17A1 requires in addition to POR also cytochrome b5 indicated by a circled b5. Similarly, hexose-6-phosphate dehydrogenase (H6PDH) is the cofactor-generating enzyme for 11 β -HSD1. The asterisk (*) indicates the 11-hydroxylation of 17OHP to 21-deoxycortisol in 21-hydroxylase deficiency. The conversion of androstenedione to testosterone is catalysed by HSD17B3 in the gonad and AKR1C3 (HSD17B5) in the adrenal. StAR, steroidogenic acute regulatory protein; CYP11A1, P450 side-chain cleavage enzyme; HSD3B2, 3 β -hydroxysteroid dehydrogenase type 2; CYP17A1, 17 α -hydroxylase; CYP21A2, 21-hydroxylase; CYP11B1, 11 β -hydroxylase; CYP11B2, aldosterone synthase; HSD17B, 17 β -hydroxysteroid dehydrogenase; CYP19A1, P450 aromatase; SRD5A2, 5 α -reductase type 2; SULF2A1, sulfotransferase 2A1; PAPSS2, 3'-phosphoadenosine 5'-phosphosulfate synthase 2; TH, tetrahydro. With permission: Krone et al. *J Steroid Biochem Mol Biol* 2010; 121:496–504.

Inhibins

Inhibins, produced by granulosa and theca cells of the ovary and Sertoli cells of the testes, play an important role in the negative feedback control of pituitary gonadotrophin secretion. The two major isoforms present in circulation are inhibin A and B. In human males, inhibin B,

a marker of Sertoli cell function, is the only biologically active form present. After an initial rise in concentration shortly after birth, inhibin B remains low until the onset of puberty, when increasing concentrations denote Sertoli cell maturation and accurately reflect spermatogenesis. In females, both inhibin A and B are produced. Inhibin B concentrations show a biphasic pattern with a

TABLE 21.2 Interpretation of serum Anti-Müllerian hormone concentrations in DSD

Serum AMH	Testicular tissue	Interpretation
Undetectable	Absent	46XX,CAH Complete gonadal dysgenesis PMDS due to AMH gene defect
Within female age-related reference range	Usually absent	46XX, CAH Dysgenetic testes or ovotestes
Below male/above female age-related reference range	Present	Dysgenetic testes Ovotestes
Within male age-related reference range	Usually normal	Non-specific XY DSD Hypogonadotrophic hypogonadism PMDS due to AMH-R defect 46XX testicular DSD Ovotestes
Above male age-related reference range	Present	AIS especially complete androgen insensitivity syndrome 5 α -reductase deficiency Testosterone biosynthetic defect Leydig cell hypoplasia

PMDS, Persistent Müllerian duct syndrome; AIS, androgen insensitivity syndrome; CAH, congenital adrenal hyperplasia.

peak at around three months, a pre-pubertal quiescent phase and a gradual rise several years prior to the onset of puberty, suggesting pre-pubertal increase in follicular activity. Inhibin A becomes detectable only in the latter stages of puberty. During the menstrual cycle, inhibin B predominates during the follicular phase reflecting recruitment of pre-antral follicles, whereas inhibin A is the major isoform produced during the luteal phase, being secreted by dominant follicles and the corpus luteum.

Original radioimmunoassays for inhibins were not able to distinguish between non-biologically active free alpha inhibin and biologically active inhibin A or B. Therefore, it was not until the development of sensitive and specific immunoassays for inhibin A and B that their potential diagnostic utility in various fields of reproductive endocrinology was truly recognized. When used in combination with standard baseline investigations of reproductive function, inhibin B has considerable potential in aiding diagnoses of disorders of pubertal development, (delay and precocity) and premature ovarian and testicular failure.

The human chorionic gonadotrophin (hCG) stimulation test

Stimulation with hCG allows the identification of functioning testicular tissue, as well as biosynthetic defects in testosterone synthesis (see Fig. 21.6). However, it is an invasive test which should only be performed as a second-line investigation after discussion with the paediatric endocrinologist in the regional DSD team. Most protocols for hCG stimulation in the UK use intramuscular hCG 1000–1500 units on three consecutive days. This can be followed by further hCG stimulation with 1500 units on two days a week for the following two weeks. In young infants and older children, three days of hCG stimulation may be sufficient. In the very young infant with an intrinsically active gonadal axis, an hCG stimulation test may not be necessary if serial blood samples show raised serum testosterone concentrations. A testosterone response to hCG may be labelled as normal if absolute testosterone

concentrations reach a level that is above the upper limit of the normal prepubertal range, or rise by more than twice the baseline value. As a minimum, other androgens that should be assessed include dihydrotestosterone (DHT) and androstenedione. For these two metabolites, the post-hCG, day four sample is more important than the pre-hCG sample. Following prolonged hCG stimulation there is no additional benefit of analysing a sample for these two metabolites on day 22. However, a day 22 sample collected for testosterone measurement can be stored and used to measure DHT or androstenedione if analysis was not achieved on day four. In the presence of a poor testosterone response following hCG stimulation, assessment of adrenal function by a standard short tetra-cosactide stimulation test should be considered. There is currently insufficient evidence to recommend that every child with XY DSD should have a tetra-cosactide stimulation test, but clinicians should be aware of the clear association between some forms of DSD and primary adrenal insufficiency. They should consider thorough assessment of adrenal function in children with diagnoses where an association has already been described and in children with any clinical suspicion of adrenal insufficiency, especially those with low steroid precursors on USP.

The role of the clinical geneticist

Establishing a specific molecular diagnosis is helpful in the clinical management of patients and in offering accurate genetic counselling for the family. However, the number of diagnostic gene tests that are available in clinically accredited DNA laboratories in the UK, or internationally, is limited and testing is costly. As developments in DNA and chromosomal analysis accelerate biomedical research, many techniques such as multiplex ligation-dependent probe amplification (MLPA) and comparative genomic hybridization (CGH) have the potential to become routine in clinical practice. Next generation DNA sequencing platforms will allow whole genome sequencing for rare diseases to become a reality at a realistic price. The clinical geneticist at the specialist DSD centre is correctly placed

TABLE 21.3 A classification system for DSD

Sex chromosome DSD	46XY DSD	46XX DSD
A: 45XO (Turner syndrome and variants)	A: Disorders of gonadal (testicular) development 1. Complete gonadal dysgenesis (Swyer syndrome) 2. Partial gonadal dysgenesis	A: Disorders of gonadal (ovarian) development 1. Ovotesticular DSD 2. Testicular DSD (e.g. SRY+, dup SOX9)
B: 47XXY (Klinefelter syndrome and variants)	B: Disorders in androgen synthesis or action 1. Androgen biosynthesis defect (e.g. 17-hydroxysteroid dehydrogenase deficiency, 5 α -reductase deficiency, StAR mutations) 2. Defect in androgen action (e.g. CAIS, PAIS)	B: Androgen excess 1. Fetal (e.g. 21-hydroxylase deficiency, 11-hydroxylase deficiency) 2. Fetoplacental (aromatase deficiency, POR deficiency)
C: 45XO/46XY (mixed gonadal dysgenesis, ovotesticular DSD)	C: Other (e.g. severe hypospadias, cloacal exstrophy)	C: Other (e.g. cloacal exstrophy, vaginal atresia, MURCS, other syndromes)
D: 46XX/46XY (chimeric, ovotesticular DSD)		

CAIS, complete androgen insensitivity syndrome; MURCS, Müllerian duct aplasia, renal aplasia, cervicothoracic somite dysplasia; PAIS, partial androgen insensitivity syndrome; POR, P450-oxidoreductase deficiency; StAR, steroidogenic acute regulatory protein.

TABLE 21.4 Characteristics of 46XX disorders of sex development due to androgen excess

	Inheritance and gene	Genitalia	Wolffian duct deriv.	Müllerian duct deriv.	Gonads	Typical signs and symptoms	Hormone profile
21-hydroxylase deficiency	Autosomal recessive, <i>CYP21A2</i>	Ambiguous	Absent	Normal	Ovary	Severe adrenal insufficiency in infancy \pm salt loss; moderate to severe androgenization at birth; \pm pigmentation	Decreased cortisol and/or mineralocorticoids Increased 17-hydroxyprogesterone, 21-deoxycortisol, androstenedione, testosterone, and/or plasma renin (activity/concentration)
11 β -hydroxylase deficiency	Autosomal recessive, <i>CYP11B1</i>	Ambiguous	Absent	Normal	Ovary	Adrenal insufficiency in infancy; moderate to severe androgenization at birth; arterial hypertension often developing at different ages	Decreased cortisol, corticosterone, aldosterone, and/or plasma renin (activity/concentration) Increased 11-deoxycortisol, 11-deoxycorticosterone, androstenedione, testosterone
3 β -hydroxysteroid dehydrogenase II deficiency	Autosomal recessive, <i>HSD3B2</i>	Commonly clitoromegaly or mild virilization, also normal	Absent	Normal	Ovary	Severe adrenal insufficiency in infancy \pm salt loss, androgenization during childhood and puberty, premature pubarche	Increased concentrations of Δ^5 C ₂₁ - and C ₁₉ -steroids, 17-hydroxypregnenolone and DHEA suppressible by dexamethasone
P450 oxidoreductase deficiency	Autosomal recessive, <i>POR</i>	Ambiguous or normal female	Absent	Normal	Ovary	Variable androgenization at birth and puberty, glucocorticoid deficiency, features of skeletal malformations. Maternal androgenization during pregnancy onset second trimester possible	Combined P450c17 and P450c21 insufficient, normal or low cortisol with poor response to ACTH stimulation, elevated 17-hydroxyprogesterone, testosterone, progesterone and corticosterone; low oestradiol
P450 aromatase deficiency	Autosomal recessive, <i>CYP19A1</i>	Ambiguous	Absent	Normal	Ovary	Delayed bone age, development of ovarian cysts during infancy, childhood and puberty. Maternal androgenization during pregnancy	High androgens in cord blood, androgens may stay elevated or normalize soon after birth

Adapted from DSD Guidelines Clinical Endocrinology 2011; 75:12–26.

the ipsilateral gonad. The minority of infants raised as girls have breast development at puberty and may also menstruate. Fertility has been reported in some cases. In contrast, individuals with 46XX testicular DSD usually have a normal male phenotype and absent Müllerian structures and are often diagnosed after karyotype analysis during investigation for infertility. In 46XX testicular DSD, about 80–90% of patients will have Y chromosomal material including a translocated *SRY* gene, which is only rarely detected in 46XX ovotesticular DSD. In other patients with 46XX testicular DSD, duplications of the *SOX9* gene and mutations of the *RSOP1* gene have been described. In patients with the suspicion of 46XX ovotesticular DSD, there is a need to assess the functional potential of testicular and ovarian tissue by a combination of biochemical testing, imaging and surgical exploration.

Disorders of Müllerian development are another group of 46XX DSD and in these patients ovarian function is usually normal but often associated with cloacal anomalies and other characteristic malformations. Although most cases of Müllerian development disorders are not associated with androgen excess, the presence of the latter, particularly in the adolescent, should alert the clinician to a possible abnormality of the *WNT4* gene.

XY DSD with low testosterone and low precursor concentrations

The differential diagnosis of 46XY DSD associated with low testosterone and precursor concentrations includes: defects occurring early in steroid synthesis (steroidogenic acute regulatory (StAR) protein, P450 side-chain cleavage (scC) enzyme/*CYP11A1*, sometimes Smith–Lemli–Optiz/*DHCR7*); LH receptor defects (*LHCGR*), and partial and complete forms of gonadal (testicular) dysgenesis (Table 21.5).

Of note, complete or partial combined 17 α -hydroxylase/17,20-lyase deficiency (*CYP17A1*) may also present with ‘low testosterone and precursor’ concentrations if DHEAS and androstenedione are the only intermediates measured. The actual diagnosis can be reached by assessment of adrenal function by measuring ACTH, tetraacetate-stimulated cortisol, plasma renin activity (PRA), 11-deoxycorticosterone (DOC), aldosterone, Δ^5 -(pregnenolone, 17OHPreg) and Δ^4 -(progesterone, 17OHP) precursors or by urine steroid analysis. Isolated 17,20-lyase deficiency and ORD might also be diagnosed by this approach. Proximal blocks (StAR, P450scC) in the pathway affect steroidogenesis in the adrenal gland as well as the developing gonad.

LH receptor defects (‘Leydig cell hypoplasia’) typically result in elevated basal LH concentrations, hyper-responsive LH to GnRH stimulation, low testosterone and precursor concentrations and impaired androgen response to hCG stimulation. No Müllerian structures will be present and adrenal function is normal. A spectrum of phenotypes has been reported including ambiguous genitalia and micropenis. In some patients, basal LH may not be elevated when the hypothalamo–pituitary–gonadal axis is quiescent (six months to late childhood).

In complete gonadal dysgenesis (Swyer syndrome), children will usually have a female phenotype with

intra-abdominal streak gonads. In some situations, ovaries or even ovaries may be found. Müllerian structures are usually present owing to impaired AMH secretion in early fetal life. Androgens and their precursor concentrations will be low, LH elevated, depending on age, and a poor or absent testosterone response to hCG stimulation is seen. Anti-Müllerian hormone concentrations will be low or undetectable. Adrenal function is usually normal, unless the underlying defect is in steroidogenic factor-1 (*SF-1*) or related adrenal or gonadal factors.

Partial gonadal (testicular) dysgenesis can present with a spectrum of phenotypes ranging from clitoromegaly, to ambiguous genitalia or severe hypospadias. Müllerian structures may or may not be present and testes of variable size and architecture are present along the path of descent. The biochemical profile is similar to complete gonadal dysgenesis, but generally less severe. If mild degrees of clitoromegaly in infancy are overlooked, a 46XY child with partial gonadal dysgenesis may first present at puberty with progressive androgenization. Genetic analysis and associated features may be useful in defining the molecular aetiology of some forms of gonadal dysgenesis.

Frasier syndrome is a rare form of gonadal dysgenesis in which there is also chronic kidney disease and a risk of gonadoblastoma arising in the streak gonads. The syndrome is caused by mutation in the *WT1* gene, located on chromosome 11p13, which encodes a zinc finger protein that functions as a transcription factor.

XY DSD with low testosterone and high steroid precursor concentrations

46XY DSD with low testosterone and increased precursor concentrations can be caused by several variants of CAH, namely 17 α -hydroxylase (*CYP17A1*) deficiency, ORD and 3 β -hydroxysteroid dehydrogenase type 2 (3 β HSD2) deficiency, caused by inactivating mutations in the corresponding genes *CYP17A1*, *POR* and *HSD3B2*. In addition, 46XY DSD with low testosterone and increased precursor concentrations, can typically be found in individuals affected by 17 β -hydroxysteroid dehydrogenase type 3 (17 β HSD3) deficiency, caused by *HSD17B3* mutations (see Table 21.5).

About 1% of patients with 46XY DSD have deficiency of *CYP17A1*. Characteristically, affected individuals present with female genitalia and low DHEA, androstenedione and testosterone concentrations. In ORD, sex steroid concentrations are characteristically low, sometimes low normal, while pregnenolone and progesterone and their metabolites accumulate, because of the combined block of *CYP21A2* and *CYP17A1* activities. There is often a relative preponderance of mineralocorticoid over glucocorticoid metabolites in affected patients but hypertension manifests only in adolescence, or later. Although baseline glucocorticoid secretion is usually sufficient, in the majority of patients, the stress response to ACTH is significantly impaired, requiring hydrocortisone cover at least in times of stress or permanent glucocorticoid replacement. 3 β -hydroxysteroid dehydrogenase type 2 deficiency (also termed Δ^4 – Δ^5 isomerase deficiency) invariably leads to glucocorticoid deficiency, as well as a variable degree of mineralocorticoid deficiency; its characteristic features are

TABLE 21.5 Characteristics of 46XY disorders of sex development due to androgen deficiency or resistance

	Inheritance and gene	Genitalia	Wolffian duct derivatives	Müllerian duct derivatives	Gonads	Typical features	Hormone profile
Leydig cell hypoplasia	Autosomal recessive, <i>LH/HCGR</i>	Female, hypospadias or micropenis	Hypoplastic	Absent	Testes	Under androgenization with variable failure of sex hormone production at puberty	Low T and DHT, elevated LH and FSH, exaggerated LH response to LHRH, poor T and DHT response to hCG stimulation
Lipoid CAH	Autosomal recessive, <i>StAR</i>	Female, rarely ambiguous or male	Hypoplastic or normal	Absent	Testes	Severe adrenal insufficiency in infancy with salt loss, failure of pubertal development, rare cases associated with isolated glucocorticoid deficiency	Usually deficient of glucocorticoids, mineralocorticoids and sex steroids
P450SCC deficiency	Autosomal recessive, <i>CYP11A1</i>	Female, rarely ambiguous or hypospadias	Hypoplastic or normal	Absent	Testes	Severe adrenal insufficiency in infancy with salt loss ranging to milder adrenal insufficiency with onset in childhood	Usually deficient of glucocorticoids, mineralocorticoids and sex steroids
3 β -hydroxysteroid dehydrogenase II deficiency	Autosomal recessive, <i>HSD3B2</i>	Ambiguous, hypospadias	Normal	Absent	Testes	Severe adrenal insufficiency in infancy \pm salt loss, poor androgenization at puberty with gynaecomastia	Increased concentrations of Δ^5 C ₂₁ - and C ₁₉ -steroids, 17-hydroxypregnenolone and DHEA suppressible by dexamethasone
Combined 17 α -hydroxylase/17,20-lyase deficiency	Autosomal recessive, <i>CYP17A1</i>	Female, ambiguous, hypospadias or micropenis	Absent or hypoplastic	Absent	Testes	Absent or poor virilization at puberty, gynaecomastia, hypertension	Decreased T, increased LH and FSH, increased plasma deoxycorticosterone, corticosterone and progesterone, decreased plasma renin activity/concentration, low renin hypertension with hypokalaemic alkalosis
Isolated 17,20-lyase deficiency	Autosomal recessive, <i>CYP17A1</i> , usually affecting key redox domains, alternatively caused by cytochrome b5 mutations (<i>CYB5</i>)	Female, ambiguous or hypospadias	Absent or hypoplastic	Absent	Testes	Absent or poor androgenization at puberty, gynaecomastia	Decreased T, DHEA, androstenedione and oestradiol, abnormal increase in plasma 17-hydroxyprogesterone and 17-hydroxypregnenolone, increased LH and FSH, increased ratio of C ₂₁ -deoxysteroids to C ₁₉ -steroids after hCG stimulation
True isolated 17,20-lyase deficiency (Cytochrome b5 deficiency)	Autosomal Recessive, <i>CYP17A1</i> , caused by missense mutations in cytochrome b5 (<i>CYB5</i>)	Female, ambiguous or hypospadias	Normal	Absent	Testes	Absent or poor androgenization at puberty	Decreased T, DHEA, normal cortisol response to ACTH stim, increased 17-OHP metabolites excretion, normal mineralocorticoid and glucocorticoid metabolites excretion. Increased methaemoglobin
P450 oxidoreductase deficiency	Autosomal recessive, <i>POR</i>	Ambiguous, hypospadias or normal male	Absent or hypoplastic	Absent	Testes	Variable androgenization at birth and puberty, glucocorticoid deficiency, features of skeletal malformations. Maternal androgenization during pregnancy onset second trimester possible	Combined P450c17 and P450c21 insufficient, normal or low cortisol with poor response to ACTH stim, elevated 17-hydroxyprogesterone, T low

TABLE 21.5 Characteristics of 46XY disorders of sex development due to androgen deficiency or resistance (Continued)

	Inheritance and gene	Genitalia	Wolffian duct derivatives	Müllerian duct derivatives	Gonads	Typical features	Hormone profile
17 β -hydroxysteroid dehydrogenase type 3 deficiency	Autosomal recessive <i>HSD17B3</i>	Female, ambiguous, blind vaginal pouch	Present	Absent	Testes	Androgenization at puberty, gynaecomastia variable	Increased plasma estrone, decreased ratio of testosterone/androstenedione and oestradiol after hCG stimulation, increased FSH and LH
5 α -reductase-2 deficiency	Autosomal recessive <i>SRD5A2</i>	Ambiguous, micropenis, hypospadias, blind vaginal pouch	Normal	Absent	Testes	Decreased facial and body hair, no temporal hair recession, prostate not palpable	Decreased ratio of 5 α :5 β C ₂₁ - and C ₁₉ -steroids in urine, increased T/DHT ratio before and after hCG stimulation, modest increase in LH, decreased conversion of T to DHT in vitro
CAIS	X-linked recessive <i>AR</i>	Female with blind vaginal pouch	Often present depending on mutation type	Absent or vestigial	Testes	Scant or absent pubic and axillary hair, breast development and female body habitus at puberty, primary amenorrhoea	Increased LH and T, increased oestradiol, FSH levels normal or slightly increased, resistance to androgenic and metabolic effects of T (may be normal in some cases)
PAIS	X-linked recessive <i>AR</i>	Ambiguous with blind vaginal pouch, isolated hypospadias, normal male with infertility (mild)	Often normal	Absent	Testes	Decreased to normal axillary and pubic hair, beard growth and body hair, gynaecomastia common at puberty	Increased LH and T, increased oestradiol, FSH levels may be normal or slightly increased, partial resistance to androgenic and metabolic effects of T

DHT, dihydrotestosterone; FSH, follicle-stimulating hormone; hCG, human chorionic gonadotrophin; LH, luteinizing hormone; T, testosterone; DHEA, dehydroepiandrosterone; ACTH, adrenocorticotropic hormone, 17-OHP, pregnanetriol; AR, androgen receptor; CAIS, complete androgen insensitivity syndrome; PAIS, partial androgen insensitivity syndrome.
Adapted from DSD Guidelines Clinical Endocrinology 2011; 75:12–26.

outlined in Table 21.5. 17 β -hydroxysteroid dehydrogenase type 3 is responsible for the conversion of androstenedione to testosterone in the gonad and deficiency has no effect on adrenal steroidogenesis. Characteristically, androstenedione concentration is increased while that of testosterone is low, particularly after hCG stimulation. However, a low testosterone to androstenedione ratio may also occur in patients with gonadal dysgenesis and the reliability of a low ratio in identifying 17 β -HSD3 deficiency is unclear. In urine, the typical finding is an increase in the androgen and androstenedione metabolites, androsterone and etiocholanolone, but it is unclear whether this applies across all age groups.

XY DSD with normal testosterone, normal precursor and low DHT concentrations

The 5 α -reductase type 2 (*SRD5A2*) isoenzyme is highly expressed in androgen sensitive tissues and converts testosterone to the more potent androgen, dihydrotestosterone (DHT) required for the development of external male genitalia. At birth, the external appearance of the genitalia of a child with *SRD5A2* deficiency can range from a completely female phenotype, through hypospadias of varying severity, to isolated micropenis. A positive family history is often present in this autosomal recessive condition. In serum, the testosterone:DHT ratio following hCG stimulation usually exceeds 30:1. In infants

over 3–6 months, the defect should be easily identifiable on a urine sample which shows a decreased ratio of 5 α :5 β -reduced C₂₁ and C₁₉ steroids and the diagnosis can therefore be made in situations where a child has had early gonadectomy. Early diagnosis of this condition is important as the affected infant may need sex reassignment if initially raised as a girl. In the child raised as a boy, application of topical DHT cream may be a way of assessing the potential of the genitalia to virilize over the longer term.

XY DSD with normal testosterone, normal precursor and normal DHT concentrations

A defect in androgen signalling is most likely to be due to dysfunction of the androgen receptor (*AR*) and mutations resulting in a complete lack of function of the *AR* cause complete androgen insensitivity syndrome (CAIS). This presents in the newborn infant as a discordance between a female phenotype and a prenatal karyotype of 46XY, as inguinal swellings in an otherwise normal girl or during a postnatal check because of a positive family history. Complete androgen insensitivity syndrome usually presents in adolescence as primary amenorrhoea with normal breast development. The presence of pubic hair is often reported in CAIS and should not be used to exclude the diagnosis. Mutations which result in some residual *AR*

function and varying degrees of androgenization cause partial androgen insensitivity syndrome (PAIS). Although children with PAIS typically have a normal testosterone and DHT response to hCG stimulation and a normal urinary steroid profile, some demonstrate a poor response to hCG stimulation. The serum AMH concentration is normal or may even be elevated. Luteinizing hormone concentrations are increased in the face of normal or elevated serum testosterone, reflecting a state of androgen resistance. A family history of X-linked inheritance is informative, although one-third of cases are the result of spontaneous new mutations.

A functional assessment of androgen sensitivity may include the clinical effect of a short course of testosterone applied to the phallus, or the effect of systemic testosterone following hCG stimulation. However, there is no consensus on the choice of androgen, dosage, method of administration, timing or duration of treatment or on the definition of a satisfactory response in the growth of the phallus. Androgen sensitivity can also be assessed by measuring the change in SHBG, an androgen-responsive protein which normally decreases in concentration following androgen exposure. This fall in SHBG is absent in CAIS, variable in PAIS and difficult to interpret in young infants who have highly variable circulating SHBG concentrations. Androgen receptor analysis may reveal a mutation in over 80% of patients with a CAIS phenotype and 30% of patients with a PAIS phenotype and androgen receptor binding studies are not necessary for routine diagnosis of AIS. A number of patients with XY DSD are loosely labelled as 'PAIS' when no conclusive biochemical or genetic abnormalities are identified in gonadal function, androgen synthesis or androgen action. The term PAIS should be reserved for those children who have XY DSD and a pathogenetic mutation in *AR*.

DISORDERS OF PUBERTY

Precocious puberty

Signs of puberty that occur before the ages of eight years in girls and nine years in boys are generally considered to indicate abnormally early puberty. The problem is more common in girls, in whom a cause is not often found. The earlier occurrence of puberty in the population has been attributed to the effects of xenoestrogenic chemicals in the environment and perhaps also to the increasing incidence of obesity. Adipocytes are a source of oestrogens through peripheral aromatization. Precocious puberty is increasingly seen in girls migrating from deprived environments for adoption in the developed world. Previous exposure to chemicals such as pesticides coupled with refeeding and catch-up growth in the new environment may be the trigger for early puberty. [Box 21.1](#) lists some of the known causes of precocious puberty in both sexes. Structural lesions in the hypothalamic-pituitary region are increasingly recognized through sensitive imaging techniques. Severe head trauma may be followed quite abruptly by the onset of puberty. It is presumed that premature activation of the GnRH pulse generator has occurred. Similar mechanisms may operate in hydrocephalus, cerebral palsy, spina bifida and intracranial infections.

BOX 21.1 Some causes of precocious puberty

Gonadotrophin dependent (central causes)

- Idiopathic (particularly girls)
- Intracranial tumours (hypothalamic hamartoma)
- Raised intracranial pressure (trauma, hydrocephalus)
- Cerebral palsy
- Cranial irradiation

Gonadotrophin independent

- McCune–Albright syndrome
- Increased peripheral sex steroids (CAH)
- Gonadal tumours
 - Male: Leydig cell adenoma, adrenal tumour
 - Female: juvenile granulosa cell tumour, ovarian cyst
- Activating LH receptor mutation (male limited)

Puberty variants

- Premature adrenarche
- Premature thelarche (female limited)
- Exposure to environmental oestrogens
- Pubertal gynecomastia

Gonadotrophin dependent puberty (central causes)

Although the cause of precocious puberty in a large majority of children remains unknown (idiopathic), some children may require a thorough endocrine evaluation of the pubertal axis as listed in [Table 21.6](#). Important causes outlined below should be considered if clinically appropriate.

Tumours in the region of the hypothalamus known to be associated with precocious puberty include hamartoma, astrocytoma, neurofibroma and, occasionally, craniopharyngioma. While the prevalence of such tumours causing precocious puberty is the same in boys and girls, a tumour is more likely to be the cause of precocious puberty in a boy. Hypothalamic-pituitary damage from irradiation to intracranial tumours causes pituitary deficiency of growth hormone and thyrotrophin releasing hormone (TRH) but, paradoxically, premature reactivation of the GnRH pulse generator and precocious puberty.

The McCune–Albright syndrome comprises the triad of café-au-lait skin pigmentation, fibrous dysplasia of bone (affecting particularly long bones and the base of

TABLE 21.6 Basic endocrine investigation of disorders of puberty

Precocious puberty	Delayed puberty
Basal LH, FSH, testosterone or oestradiol	Basal LH, FSH, testosterone or oestradiol
Acute GnRH stimulation test	Acute GnRH stimulation test
Thyroid hormone profile (TSH & FT4)	Thyroid function tests (TSH & FT4)
Prolactin	Peripheral karyotype (females)
β-hCG	Possibly hCG stimulation test (males)
Skeletal age (bone age)	Skeletal age (bone age)

the skull) and precocious puberty. In girls, the latter is typically gonadotrophin independent, as shown by a prepubertal LH and FSH response to acute GnRH stimulation. Autonomously functioning multiple ovarian cysts occur, as well as evidence of dysfunction in other endocrine glands such as the thyroid (thyrotoxicosis), adrenal (Cushing syndrome), pituitary (gigantism and hyperprolactinaemia) and parathyroid (hyperparathyroidism). The condition may present in early childhood and progress to gonadotrophin-dependent ovarian hyperfunction and the consequent problems of menses in a young child. The McCune–Albright syndrome is less common in boys and may manifest with asymmetrical enlargement of the testes as well as the development of secondary sexual characteristics. Somatic activating mutations of the G-protein α subunit ($GS\alpha$) component of the guanine nucleotide binding protein, alpha stimulating (GNAS) complex that couples seven transmembrane receptors to the cAMP signal transduction pathway, cause the widespread distribution of tissue abnormalities in the syndrome. The predominant mutation is substitution of an arginine at codon 201 in the $GS\alpha$ gene. Heterozygous inactivating $GS\alpha$ mutations cause Albright hereditary osteodystrophy, a syndrome characterized by short stature, obesity, short metacarpals, subcutaneous calcification and behavioural problems.

Autonomously hyperfunctioning testes, independent of gonadotrophic stimulation, also cause precocious puberty in males, owing to Leydig cell hyperplasia. Familial forms are recognized that present with signs of virilization, but with testes inappropriately small for the advanced stage of puberty. The onset of precocious puberty often occurs in early childhood and may even occur in infancy. Testosterone concentrations are increased, while basal and GnRH-stimulated concentrations of gonadotrophins are suppressed. Histological examination of the testes shows Leydig cell hyperplasia and spermatogenesis. This form of familial, male-limited precocious puberty is autosomal dominant and is caused by an activating mutation in the G-protein-LH-coupled receptor. Somatic LH receptor mutations are also found in males with Leydig cell adenomas without a history of precocious puberty, but a Leydig cell adenoma rarely occurs in the precocious puberty phenotype. This is at variance with the phenotype associated with activating mutations in the TSH receptor where nodular thyroid hyperplasia is often an associated histological feature of this form of hyperthyroidism. Treatment for male-limited, familial precocious puberty comprises an anti-androgen agent, such as spironolactone, together with an aromatase enzyme inhibitor such as testolactone or anastrozole, but the long-term therapeutic benefits of this approach are unclear. Use of a GnRH analogue is not appropriate initially as gonadotrophin concentrations are already suppressed. However, the precocious puberty may become gonadotrophic dependent at a later stage, when a long-acting GnRH analogue can be added to the treatment regimen.

Variants of early puberty

The most frequent variant of early puberty is premature thelarche, or isolated early breast development. The usual age of onset is 1–3 years and occasionally breast

development may have persisted from the neonatal period. The breast enlargement may be unilateral or bilateral and typically tends to wax and wane over time. No other signs of puberty occur and the rate of linear growth is normal for chronological age. Serum oestradiol and gonadotrophin concentrations remain prepubertal, although most immunoassays are insufficiently sensitive to detect fluctuations in oestradiol concentrations in prepubertal girls. It has been proposed that breast enlargement is the result of enhanced tissue responsiveness to a transient increase in circulating oestradiol. Increased concentrations of SHBG have been detected in some girls with premature thelarche, thereby resulting in an increase in the ratio of free oestradiol to testosterone concentrations. This may explain the breast development, even though total oestradiol concentration is not increased. The source of oestrogen is ovarian, and cysts of the ovary are sometimes detected on ultrasound. While population studies imply the possibility of exposure to chemicals in the environment acting as xenoestrogens, it is difficult to prove such exposure in individual patients. The natural history of premature thelarche is the onset of puberty and menarche at a normal age, as well as normal reproductive potential. A form of thelarche variant is also recognized that is associated with some increase in growth rate and skeletal maturation and a slightly exaggerated LH and FSH response to acute GnRH stimulation.

Isolated recurrent vaginal bleeding can sometimes occur in girls before puberty. It is important to exclude local lesions of the genital tract, self-harm by the child or child abuse. Generally, it is associated with increased FSH concentrations that produce sufficient oestradiol to stimulate the endometrium and induce a subsequent withdrawal bleed. Sometimes, ovarian cysts are found on ultrasound. This form of premature isolated menarche is generally self-limiting and has no adverse consequences for subsequent puberty and reproductive capacity.

Premature adrenarche is the term applied to the early development of pubic hair without any other signs of puberty. It is more common in girls, occurring usually between six and eight years of age. The cause is an increase in adrenal androgen production associated with elevated concentrations of DHEA and its sulphate, as well as androstenedione. It is partly explained by a change in steroidogenesis regulated by the orphan nuclear receptor, NGF1B. Premature adrenarche is generally followed by a normal progression through puberty. There is some evidence that girls with premature adrenarche have a higher risk of later developing polycystic ovarian syndrome.

Delayed puberty

The wide range for the age of onset of puberty in normal children has previously been emphasized. Puberty is regarded as delayed if physical signs have not started by 13 years in girls and 14 years in boys. [Box 21.2](#) lists some of the causes of delayed puberty in both sexes. The list is not an exhaustive one but encompasses the majority of important causes. During the clinical assessment it is important to exclude a chronic disease that may manifest solely as slow growth and delayed puberty, examples include cystic fibrosis, inflammatory bowel disease and anorexia nervosa.

BOX 21.2 Some causes of delayed puberty

- Constitutional delay in development
- Hypogonadotrophic hypogonadism
 - CNS tumours (craniopharyngioma, germinoma)
 - Kallmann syndrome
 - Prader–Willi syndrome
 - *GPR54* mutation
 - Inactivating GnRH receptor mutation
 - Congenital adrenal hypoplasia (associated with *DAX1* mutation)
 - Leptin deficiency
- Hypothyroidism
- Cushing disease
- Hyperprolactinaemia
 - Primary gonadal failure
 - Turner syndrome
 - Gonadal dysgenesis (XX and XY)
 - Chemotherapy (e.g. cyclophosphamide)
 - Radiation therapy
- Chronic systemic dysfunction
 - Inflammatory bowel disease
 - Coeliac disease
 - Chronic kidney disease
 - Psychosocial deprivation
 - Anorexia nervosa
 - Intense exercise

Delayed growth and puberty

Delayed puberty presents as a clinical problem far more often in boys than in girls. This may be an ascertainment bias, perhaps because of the greater social pressures of delayed puberty (and hence short stature) in a boy than a girl. Most delayed puberty just represents the extreme lower end of the physiological range for pubertal events. Since it is inevitably accompanied by short stature due to the delayed appearance of the pubertal growth spurt, this common cause is called constitutional delay in growth and development. There is often a family history of delayed puberty. The natural outcome is for puberty to develop spontaneously (according to skeletal maturation as assessed by bone age examination rather than chronological age) and lead to normal, albeit delayed, final height according to genetic potential. Sometimes, low-dose sex steroids are given to prime the pubertal growth spurt. Final height may be increased in boys with constitutional delay by treatment with aromatase inhibitors during adolescence, but there are concerns that such treatment may have undesired effects on bone health at this critical period of skeletal development.

Delayed puberty is a disturbance in the tempo of growth and pubertal maturation associated with a delay in activation of the hypothalamic GnRH pulse generator. Random and GnRH-stimulated plasma gonadotrophin concentrations are low and appropriate for physiological development, as are those of gonadal steroids. A similar endocrine profile is found in delayed puberty secondary to hypogonadotrophic hypogonadism in which spontaneous onset of puberty does not occur. It can be difficult to distinguish these two variants, even when short-term pulsatile GnRH therapy is used in the prepubertal state. Spontaneous onset of puberty is the ultimate decider.

Familial clustering may suggest rare causes such as mutations affecting the GnRH receptor gene or the *GPR54* gene. No mutation has yet been found involving the *GnRH* gene and delayed puberty. Short stature in delayed puberty is partly the result of an insufficient production of growth hormone (GH) reflected in decreased pulse amplitude. This, in turn, results in less generation of insulin-like growth factor 1, which may be relevant for gonadal function. Puberty is a complex interplay of sex hormones, GH and growth factors activated as a result of polygenic and environmental factors. While the discovery of *GPR54* has provided a molecular explanation for a rare form of hypogonadotrophic hypogonadism, no *GPR54* mutations have been reported in constitutional delayed puberty.

Organic causes affecting the hypothalamic-pituitary area result in hypogonadotrophic hypogonadism and delayed puberty. Numerous examples exist such as craniopharyngioma, optic glioma, germinoma, astrocytoma, head trauma (which may also lead to early puberty), effects of irradiation, infiltrative diseases and post-infectious lesions. An organic cause should be considered when puberty has started, but then becomes arrested in progress.

Hypogonadotrophic hypogonadism

Syndromes associated with gonadotrophin deficiency include Prader–Willi, Lawrence–Moon–Biedl and Kallman syndromes. Failure of migration of the GnRH neurons from the olfactory placode to the medial basal hypothalamus causes Kallman syndrome and explains the association with anosmia. An X-linked form of Kallman syndrome is caused by mutations in the *KALI* gene, which encodes for the anosmin-1 protein. An autosomal dominant form is caused by heterozygous mutations in the *FGFR1* gene (fibroblast growth factor receptor-1). Inactivating mutations in the *GPR54* gene cause isolated hypogonadotrophic hypogonadism.

Isolated gonadotrophin deficiency may occur in association with congenital adrenal *hypoplasia*. Affected individuals present in infancy with a salt-losing crisis and masquerade as if the diagnosis were congenital adrenal *hyperplasia*. However, adrenal steroid concentrations are low and unresponsive to ACTH stimulation. The diagnosis may be suspected before birth because of low maternal oestriol excretion resulting from decreased steroid substrates from the affected fetal adrenals. The disorder is X-linked; affected boys fail to enter puberty spontaneously because of an associated gonadotrophin deficiency. Adult males have azoospermia and a poor response to pulsatile GnRH treatment. Mutations in *DAX-1*, a gene on Xp21.3–21.2 that encodes an orphan nuclear hormone receptor, are described in patients with X-linked congenital adrenal hypoplasia and hypogonadotrophic hypogonadism. A variant of the syndrome may present for the first time in adulthood with mild adrenal insufficiency and partial hypogonadism. Another orphan nuclear receptor involved in both adrenal and hypothalamic development is steroidogenic factor-1 (SF-1). Rare mutations cause adrenal failure and generally XY sex reversal.

Primary hypogonadism

The other major groups of disorders that manifest with delayed puberty are those associated with primary hypogonadism. In these, gonadotrophin concentrations are elevated. The chromosome aneuploidies, Klinefelter syndrome (47XXY) and Turner syndrome (45XO and variants), are important causes of delayed puberty. Turner syndrome has an incidence of about 1 in 2000 live-born girls. It is estimated that 1–2% of conceptuses have the syndrome with only 1% of these not aborted spontaneously. Two constant clinical features of this syndrome are short stature and failure to enter puberty spontaneously because of premature ovarian failure. The former feature dictates that a peripheral karyotype should be performed on any short girl in whom there is no readily apparent cause for the growth failure.

The typical physical features of Turner syndrome may not always be present. They include a short, broad neck which may be webbed, a low hairline, low-set ears, high-arched palate, ptosis, hypoplastic nails, short metacarpals, cubitus valgus (increased carrying angle), pigmented naevi and cardiovascular anomalies. Affected girls are often small at birth and have peripheral lymphoedema. This usually is an alert to making the diagnosis. The results of a pilot study using FSH measurement in blood spots for newborn screening were not sufficiently specific. Pyrosequencing assays for quantitative genotyping of informative single nucleotide polymorphism (SNP) markers spanning the X chromosome in a newborn screening programme reliably detect complete or partial X deletions and mosaicisms. Growth failure is invariably present by 2–3 years of age. Ovarian failure is a progressive process starting in late gestation and appears to be an acceleration of the diminution in the number of primary oocytes that occurs between fetal and postnatal life in the normal female. Generally, streak gonads are the only remnants of ovaries, although 5–10% of Turner syndrome girls do have spontaneous puberty and menstruate. A small number are able to become pregnant, but this is associated with a high rate of miscarriage, stillbirths and congenital malformations. Management of Turner patients requires oestrogen replacement at the time of expected puberty and the early use of GH to enhance adult height. Numerous studies confirm that GH can significantly increase final height. Chromosomal mosaicism in the form of 45XO/46XY may show some features of the Turner phenotype, such as short stature, but the presentation may be at birth because of abnormal genitalia. Adult patients with Turner syndrome need regular surveillance of cardiovascular parameters, metabolic abnormalities such as insulin resistance, renal disorders and sensorineural hearing deficits. Turner syndrome represents one example of spontaneous premature ovarian failure. In addition, there are other causes of acquired as well as genetic causes of ovarian failure.

Klinefelter syndrome affects about 1 in 1000 males and is characterized by tall stature, small firm testes, gynaecomastia and infertility in adults. These men are also hypogonadal and, although they enter puberty normally, they undergo pubertal arrest and progress through puberty does not occur as expected. Serum

testosterone concentrations increase appropriately according to the stages of puberty, while a characteristic elevation of gonadotrophin concentrations is displayed from the age of 13 years. Fertility may be achievable using testicular extraction for intracytoplasmic sperm injection. Other causes of primary hypogonadism in the male include testosterone biosynthetic defects and the effects of chemotherapy in conditions such as leukaemia. Anorchia (absence of both testes at birth) is relatively common and is sometimes referred to as the vanishing testis syndrome. Genital development in affected boys is otherwise normal, which implies that testicular function (including testosterone and AMH production) was also normal in early gestation. The cause of testicular regression is not known but is postulated to be the result of intrauterine testicular torsion. The presence or absence of testicular tissue after birth can be assessed by a prolonged hCG stimulation test to determine any response in serum testosterone. Serum AMH is also a useful marker that avoids the necessity of undertaking an hCG stimulation test. At the expected age of puberty, there is an increase in plasma gonadotrophins.

Further reading

Disorders of sex development

- Achermann JC, Ozisik G, Meeks JJ et al. Genetic causes of human reproductive disease. *J Clin Endocrinol Metab* 2002;87:2447–54.
- Ahmed SF, Achermann JC, Arlt W et al. UK guidance on the initial evaluation of an infant or an adolescent with a suspected disorder of sex development. *Clin Endocrinol (Oxf)* 2011;75:12–26.
- Ahmed SF, Hughes IA. The genetics of male undermasculinisation. *Clin Endocrinol (Oxf)* 2002;56:1–18.
- Broekmans FJ, Visser JA, Laven JS et al. Anti-Müllerian hormone and ovarian dysfunction. *Trends Endocrinol Metab* 2008;19:340–7.
- Cordts EB, Christofolini DM, Dos Santos AA et al. Genetic aspects of premature ovarian failure: a literature review. *Arch Gynecol Obstet* 2011a;283:635–43.
- Eggers S, Sinclair A. Mammalian sex determination – insights from humans and mice. *Chromosome Res* 2012;20:215–38.
- Federman DD. The biology of human sex differences. *N Engl J Med* 2006;354:1507–14.
- Grinspon R, Chemes H, Rey RA. Decline in serum anti-Müllerian hormone due to androgen action in early puberty in males. *Fertil Steril* 2012;98:e23.
- Hughes IA, Houk C, Ahmed SF et al. LWPES Consensus Group, ESPE Consensus Group. Consensus statement on management of intersex disorders. *Arch Dis Child* 2006;91:554–63.
- Ogilvy-Stuart AL, Brain CE. Early assessment of ambiguous genitalia. *Arch Dis Child* 2004;89:401–7.
- Speiser PW, White PC. Congenital adrenal hyperplasia. *N Engl J Med* 2003;349:776–88.
- Styne DM, Grumbach MM. Puberty: ontogeny, neuroendocrinology physiology and disorders. In: Kronenberg HM, Melmed S, Polonsky KS et al. editors. *Williams textbook of endocrinology*. 11th ed. Philadelphia: Saunders; 2007. pp. 969–1166.

Puberty

- Achermann JC, Hughes IA. Disorders of sex development. In: Kronenberg HM, Melmed S, Polonsky KS et al. editors. *Williams textbook of endocrinology*. 11th ed. Philadelphia: Saunders; 2007. pp. 783–848.
- Aubert ML, Pralong FP. Puberty: a sensor of genetic and environmental interactions throughout development. *Mol Cell Endocrinol* 2006;254/255:1–233 [Special Issue].
- Cordts EB, Christofolini DM, Dos Santos AA et al. Genetic aspects of premature ovarian failure: a literature review. *Arch Gynecol Obstet* 2011b;283:635–43.

Dattani MT, Hindmarsh PC. Normal puberty and abnormal puberty. In: Brook CG, Clayton PE, Brown RS, editors. *Clinical paediatric endocrinology*. 5th ed. Oxford: Blackwell; 2005. pp. 90–112.

Parent A-S, Teilmann GT, Juul A et al. The timing of normal puberty and the age limits of sexual precocity: variations around

the world, secular trends, and changes after migration. *Endocr Rev* 2003;24:668–93.

Slyper AH. The pubertal timing controversy in the USA, and a review of possible causative factors for the advances in timing of onset of puberty. *Clin Endocrinol (Oxf)* 2006;65:1–8.

Reproductive function in the female

Leslie D. Ross

CHAPTER OUTLINE

INTRODUCTION 433

PHYSIOLOGY 433

- The ovaries 433
- Plasma concentrations of reproductive hormones 435
- Uterine changes 436
- Conception 436

HORMONES REGULATING REPRODUCTIVE FUNCTION 436

- Follicle stimulating hormone 436
- Luteinizing hormone 436
- Human chorionic gonadotrophin 436
- Inhibin and activin 436
- Prolactin 437
- Anti-Müllerian hormone 437

REPRODUCTIVE STEROID HORMONES 437

- Structure 437
- Biosynthetic enzymes 437
- Steroid secretion through the menstrual cycle 438
- Steroid hormone transport and metabolism 438
- Actions of gonadal steroid hormones 439

OLIGO- AND AMENORRHOEA 439

INFERTILITY 440

HIRSUTISM AND VIRILISM 441

PREGNANCY 442

- Introduction 442
- Biochemical diagnosis of pregnancy 442
- Diagnosis of ectopic pregnancy 442
- Biochemical monitoring of pregnancy 443
- Screening for fetal malformation 443
- Fetal tissue sampling techniques 444
- Chorionic villus sampling 444
- Monitoring of maternal and fetal well-being 445
- Intrapartum fetal monitoring 445
- Biochemical changes during pregnancy 445
- Labour 446

ORAL CONTRACEPTION AND HORMONE REPLACEMENT THERAPY 447

- Introduction 447
- Metabolic effects of oestrogens 448
- Metabolic effects of progestogens 448
- Metabolic effects of contraceptives 448
- Hormone replacement therapy 449

APPENDIX 450

INTRODUCTION

In both normal males and females, the gonads produce steroid hormones that affect secondary sexual characteristics, the functioning of the reproductive tract and sexual behaviour. Production of gametes and hormones by the gonads is under the control of pituitary glycoprotein hormones. The hypothalamo–pituitary–gonadal axis appears to be universal throughout the vertebrates: small amounts of releasing factors from the hypothalamus elicit the release of larger amounts of glycoprotein gonadotrophins and yet larger increases in gonadal steroids. These steroids in turn influence the rate at which they themselves are produced (feedback control).

PHYSIOLOGY

The ovaries

The human ovaries produce female gametes and steroid sex hormones. Both functions depend largely on the monthly growth and rupture of (usually) a single ovarian follicle.

At birth, some 10^6 immature germ cells are present in the ovaries as primary oocytes, arrested between prophase and metaphase of the first meiotic division. Each primary oocyte is surrounded by a layer of epithelial cells, the whole being known as a primordial follicle (Fig. 22.1). Primary oocytes do not complete meiosis during childhood: in fact the majority of them degenerate.

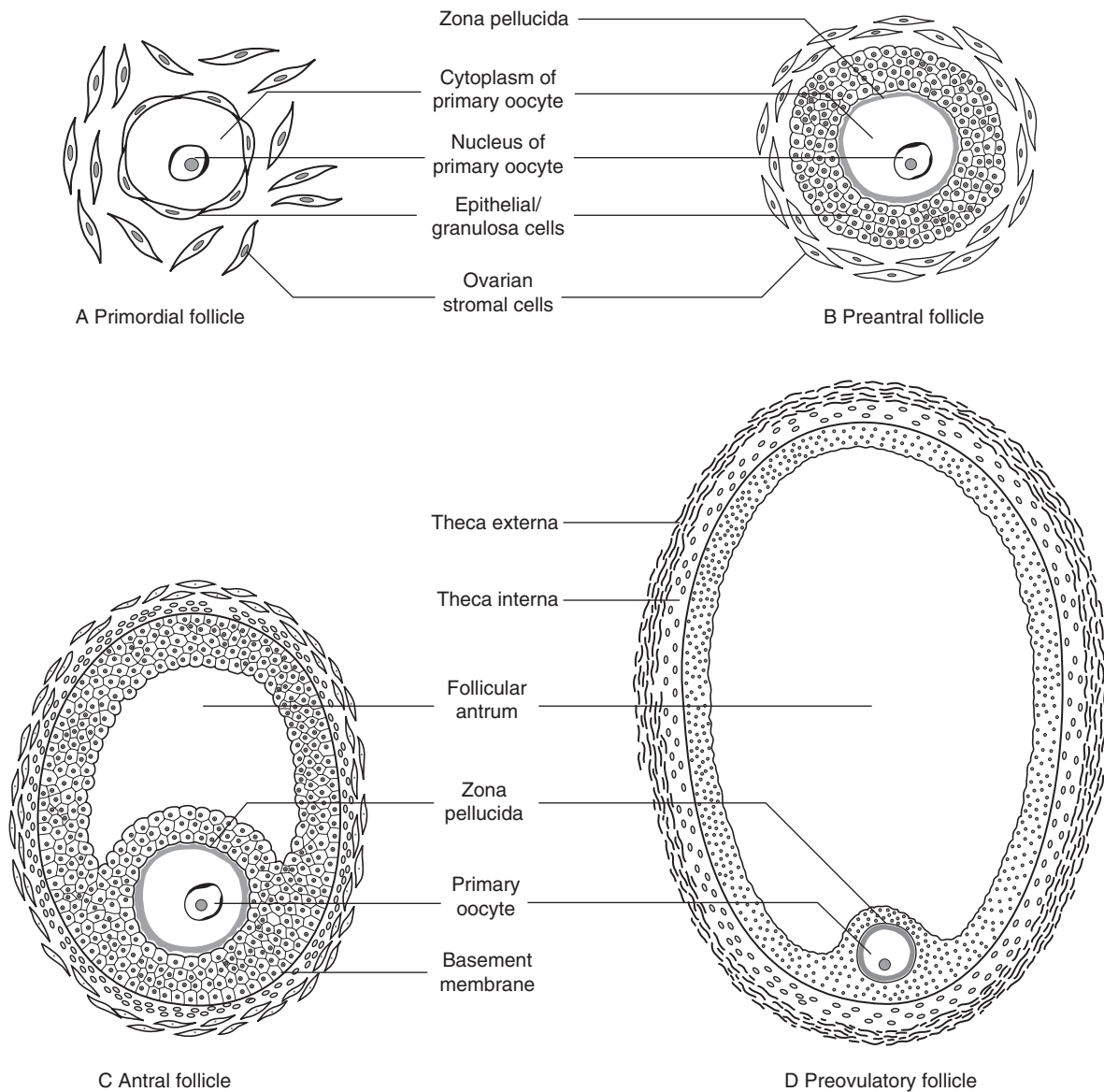


FIGURE 22.1 ■ Maturation of ovarian follicles.

During childhood, the ovaries remain inactive, but at puberty, a monthly ovarian cycle is established through the interaction of the hypothalamus and pituitary with ovarian follicles and manifests itself by the onset of menstruation.

At puberty, the pulsatile release of gonadotrophin-releasing hormone (GnRH) from the hypothalamus stimulates the pituitary production of gonadotrophins. Growth hormone (GH) pulses released by the pituitary also increase in amplitude. This amplification of GH secretion may be regulated by the pubertal increase in the production of androgenic and oestrogenic hormones. These sex steroids stimulate skeletal growth and sexual maturation, augmenting the role of GH in promoting somatic growth and development.

The human menstrual cycle is 23–39 days long. By definition, day 1 of the cycle is the first day of menstrual bleeding. The cycle is divided into follicular (or proliferative) and luteal (or secretory) phases by the event of

ovulation. In cycles of different length, it is the duration of the follicular phase that varies; the length of the luteal phase is remarkably constant at 13–15 days.

On day one of the cycle, several antral follicles, which are 2–9 mm in diameter, are present in both ovaries. These consist largely of a fluid-filled antrum and can be visualized by ultrasonography (Fig. 22.2A). Each antral follicle has developed from a primordial follicle by proliferation of the epithelial granulosa cells and by the appearance and coalescence of fluid-filled spaces among them (see Fig. 22.1). The earlier stages of follicle development do not require gonadotrophin stimulation, but the 'recruitment' of antral follicles from the preantral pool appears to be follicle stimulating hormone (FSH) and anti-Müllerian hormone (AMH) dependent.

Anti-Müllerian hormone is now considered to be the main regulator of early follicular recruitment from the primordial pool. It is produced by small enlarging follicles rather than the primordial follicles themselves. Plasma

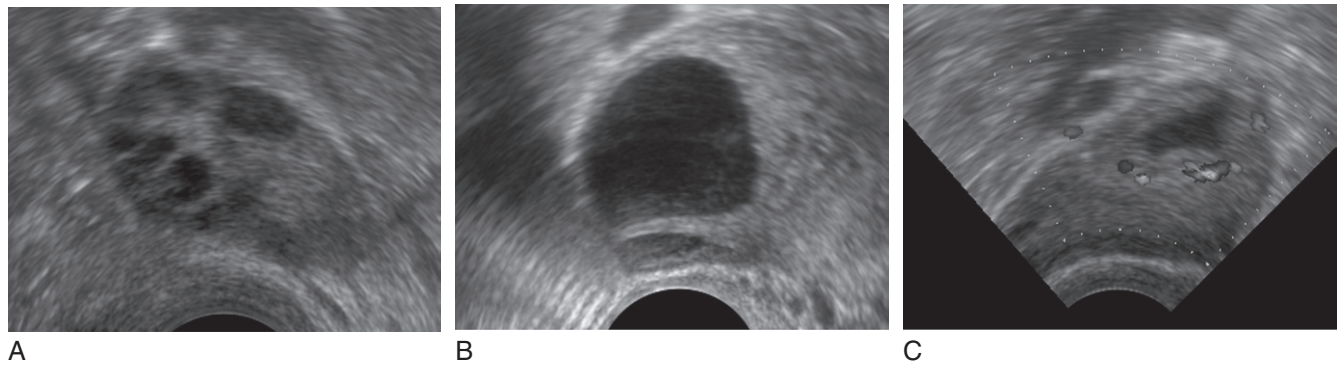


FIGURE 22.2 ■ Vaginal ultrasonographic images of the ovary. (A) Polycystic ovary showing multiple antral type peripheral follicles; (B) Pre-ovulatory follicle 20 mm; (C) Corpus luteum with Doppler showing surrounding neovascularity. Vessels show as prominent echoes below the darker collapsed follicle.

concentrations of AMH do, however, appear to reflect primordial follicle numbers in humans and rodents.

Through a poorly understood process of selection, one of the apparently identical antral follicles present on day 1 becomes dominant, while the others degenerate. The dominant follicle grows rapidly in the late follicular phase, reaching a maximum diameter of approximately 20 mm (Fig. 22.2B). Outside the basement membrane of the granulosa cell layer, the wall of the dominant follicle consists of the theca interna and theca externa, which have developed from ovarian stromal cells. At ovulation, the follicle collapses, releasing its fluid and the oocyte, which by now has completed the first meiotic division. The follicle subsequently refills with fluid, and blood vessels penetrate the basement membrane and vascularize the granulosa cells for the first time; further proliferation of the granulosa cells transforms what was the follicle into the corpus luteum (Fig. 22.2C). The corpus luteum has a limited life span: in the absence of conception it involutes, another dominant follicle develops and the cycle repeats itself.

Insulin-like growth factor-1 has actions that regulate sex steroids through the control of the plasma concentration of sex hormone binding globulin (SHBG). Although the precise mechanism is unclear, increasing concentrations of GH are believed to exert some effects on circulating insulin concentrations. Growth hormone induces peripheral insulin resistance which leads to compensatory increases in insulin secretion and protein anabolism. Insulin regulates hepatic insulin-like growth factor 1 (IGF-1) through its effects on insulin-like binding protein 1 (IGFBP-1).

Obesity is associated with increased plasma concentrations of insulin. Therefore, if excessive nutrition intake persists during childhood, it is possible that hyperinsulinaemia may lead to lower concentrations of IGFBP-1 and reduced plasma SHBG concentrations, enhancing IGF-1 and sex steroid bioavailability.

Plasma concentrations of reproductive hormones

Figure 22.3 illustrates the fluctuations in plasma concentrations of reproductive hormones through a typical menstrual cycle. Changes in the pulsatile release of gonadotrophins in the late luteal phase and early follicular phase

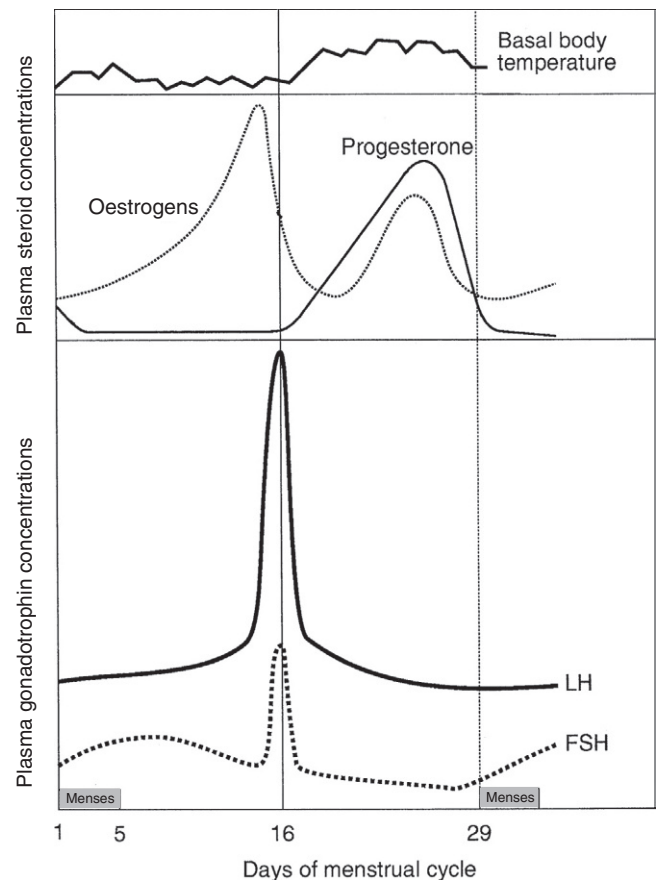


FIGURE 22.3 ■ Hormone fluctuations during the human menstrual cycle.

appear to bring about the growth of a group of antral follicles, one of which (the dominant follicle) enlarges greatly and secretes increasing amounts of oestrogens (mainly 17 β -oestradiol). Although oestrogens generally exert a negative feedback effect on gonadotrophin concentrations, high, rising oestrogen concentrations in the late follicular phase feed back positively on the hypothalamic–pituitary axis, causing a massive release of luteinizing hormone (LH) (the LH ‘surge’) and a smaller release of FSH. The LH surge triggers the resumption of meiosis and follicle rupture with oocyte release. Oestrogen concentrations fall transiently prior to ovulation. The corpus luteum secretes

both oestrogens and progesterone; plasma concentrations of both hormones peak in the mid-luteal phase falling in the late luteal phase as, in the absence of conception, the corpus luteum involutes. It is likely that ovarian proteins such as inhibin and activin also play a role in regulating follicle development and steroid synthesis. These proteins and other structurally similar 'growth factors' may exert important paracrine effects within the ovary.

Uterine changes

Rising oestrogen concentrations in the follicular phase induce proliferative changes in the endometrium, causing it to thicken. Oestrogens and progesterone (from the corpus luteum) induce secretory changes in the endometrium, preparing it for pregnancy. Once deprived of hormonal support from the corpus luteum, the endometrium is shed and another cycle begins.

Conception

Fertilization usually takes place in the ampullary part of the Fallopian tube, close to the ovary. Implantation occurs 6–7 days after fertilization. Human chorionic gonadotrophin (hCG) production from the conceptus acts to prolong the life of the corpus luteum so that it continues to produce progesterone, thus maintaining the endometrium.

HORMONES REGULATING REPRODUCTIVE FUNCTION

Gonadotrophin releasing hormone is a decapeptide secreted by neurons in the median eminence of the hypothalamus. It is carried in the portal hypophysial vessels from the hypothalamus to the pituitary, where it stimulates the release of LH and FSH. Concentrations of GnRH in the general circulation are very low but are assumed to be reflected by LH concentrations. The release of LH and FSH are pulsatile: the inter-pulse interval decreases from 90 to 60 min during the follicular phase until ovulation, then lengthens to once or twice a day in the late-luteal phase.

Pulsatile secretion of GnRH is essential to maintain normal function, as continuous infusion causes downregulation of receptors and amenorrhoea. Administration of GnRH, modified to prolong its action, is used to induce therapeutic hypogonadism.

Follicle stimulating hormone, LH and hCG are structurally similar glycoproteins. Each consists of an α - and a β -subunit, which are associated by non-covalent interactions. The α -subunits of FSH, LH and hCG (and of thyroid stimulating hormone, TSH) are identical, while the β -subunits differ, conferring hormonal specificity. The α -subunit consists of 92 amino acids. Gonadotrophins are largely metabolized in the liver; 10–15% is excreted unchanged in the urine.

Follicle stimulating hormone

The β -subunit of FSH contains 118 amino acids. Carbohydrate constitutes 22–25% of the dry weight of

FSH and there are four branched, mannose-rich oligosaccharide side chains covalently bound to asparagine residues, two on the α -subunit and two on the β -subunit. The biological role of the carbohydrate is partly to protect the hormone against degradation. Follicle stimulating hormone is synthesized and secreted by gonadotrophic cells of the anterior pituitary, which also synthesize LH.

The primary target of FSH is the granulosa cell of the ovarian follicle. Follicle stimulating hormone stimulates differentiation of immature granulosa cells, induces cytochrome P450-dependent aromatase, induces LH receptors on granulosa cells and increases the binding of FSH by granulosa cells. The circulating half-life of FSH is ~4h. Oestradiol amplifies the actions of FSH on granulosa cells.

Luteinizing hormone

The β -subunit of LH consists of 92 amino acids. Luteinizing hormone has three carbohydrate side chains, carbohydrate accounting for approximately 15% of the dry weight of the hormone. The circulatory half-life of LH is short (20 min), contributing to the pronounced pulsatility of plasma LH concentrations.

Ovarian LH receptors are found on theca cells and on mature granulosa cells. Luteinizing hormone stimulates production of androgens by the theca cells, and of oestradiol and progesterone by mature granulosa cells and corpus luteal cells. Luteinizing hormone regulates steroid biosynthesis by influencing uptake and side chain cleavage of cholesterol.

Human chorionic gonadotrophin

The β -subunit of hCG consists of 145 amino acids. The hormone is synthesized by the syncytiotrophoblastic cells of the placenta. It has a relatively long plasma half-life (24–36 h).

Human chorionic gonadotrophin interacts with the same receptors on luteal cells as does LH. It acts in the first trimester of pregnancy to maintain the corpus luteum and its ability to secrete progesterone, which supports the endometrium. After the first trimester, the placenta takes over the synthesis of progesterone from the corpus luteum. Human chorionic gonadotrophin is detectable in plasma eight days after conception and its concentration peaks around the tenth week of pregnancy.

Human chorionic gonadotrophin also stimulates testosterone synthesis in the testes of male fetuses, providing stimulation for male sexual differentiation.

Inhibin and activin

Inhibin is a glycoprotein that has been isolated from follicular fluid in two forms comprising a common α -subunit and one of two β -subunits, β_A (in inhibin A) and β_B (in inhibin B). Follicular fluid also contains two dimers of the inhibin β -subunit, which are called activin: activin A is the homodimer of the β_A -subunit, activin A-B is the $\beta_A\beta_B$ heterodimer.

Inhibin A and inhibin B have different patterns of circulation during the two phases of the ovarian cycle and play different physiological roles during follicular recruitment, maturation and ovulation. In the luteal phase,

inhibin A suppresses FSH secretion. The concentration of inhibin A then decreases significantly as the corpus luteum involutes.

Development of the dominant follicle is characterized by secretion of increasingly large amounts of oestradiol and inhibin A into the circulation. There is evidence that the maintenance of the follicle is affected by intraovarian signalling, with inhibins and activins acting as important paracrine messengers.

During the menstrual cycle, plasma activin A concentrations vary in a biphasic manner, with the highest values occurring at mid-cycle during the luteo-follicular transition and nadirs occurring in the mid-follicular and mid-luteal phases.

Prolactin

Prolactin is a single chain polypeptide comprising 199 amino acids. It shares a high degree of homology with growth hormone and placental lactogen. Its most potent biological form (80–90%) is monomeric; 8–20% is dimeric and 1–5% is macroprolactin. The latter is a complex of monomeric prolactin and IgG and is immunoreactive (i.e. detected in most assay systems) but biologically relatively inactive.

Prolactin is synthesized by lactotroph cells of the anterior pituitary. Dopamine is the principal negative modulator of prolactin secretion; oestrogens and thyroid hormone releasing hormone increase prolactin release.

Prolactin has a variety of actions in vertebrates, playing roles in processes as diverse as osmoregulation and metamorphosis. It also has a direct but variable effect on follicular development and function. Prolactin is luteotrophic in some mammals but not in man. The only definite role of prolactin in women is the postpartum initiation and maintenance of milk production.

Macroprolactinaemia is a relatively common phenomenon in patients with hyperprolactinaemia in clinical practice; to ensure appropriate detection and management, all patients found to have significant hyperprolactinaemia should be screened for macroprolactinaemia (see Chapter 18).

Anti-Müllerian hormone

Anti-Müllerian hormone is a glycoprotein also known as Müllerian inhibitory factor (MIF) (see Chapter 21 for its role in normal and disordered sex development). It belongs to the transforming growth factor b family and is encoded on the short arm of chromosome 19. It acts through two receptors (AMHR1 and 2), which are present on the target organs in the gonads and Müllerian ducts.

Measurement of AMH is used clinically for assessment of 'ovarian reserve' and likely response to ovarian stimulation for subfertility. It has largely superseded measurement of basal FSH and inhibin B as a marker of follicular potential, particularly as the concentrations do not change with the menstrual cycle. It can also be used as a tumour marker for granulosa cell malignancies and appears to be an excellent marker for polycystic ovarian syndrome as these patients have significantly elevated concentrations.

REPRODUCTIVE STEROID HORMONES

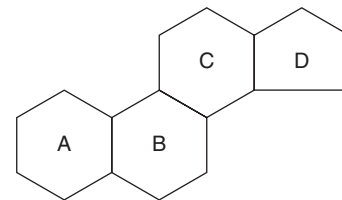
Structure

Ovarian steroid hormones are derivatives of cholesterol and contain the cyclopentanoperhydrophenanthrene nucleus (Fig. 22.4). Oestrogens (C-18 steroids), androgens (C-19 steroids) and progestogens (C-21 steroids) contain 18, 19 and 21 carbon atoms, respectively. These compounds and the pathways involved in their biosynthesis are illustrated in Figure 22.5.

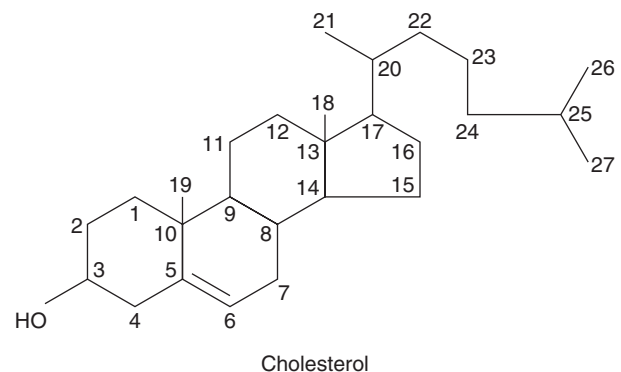
Biosynthetic enzymes

The numbers below refer to the enzymes and pathways labelled in Figure 22.5.

1. *Cholesterol side chain cleavage*. Three separate reactions are involved: 20 α -hydroxylation, 22-hydroxylation and splitting of the C20–22 carbon bond. This is the point where LH primarily regulates ovarian steroidogenesis. The product of the process, which occurs in a mitochondrial complex containing the cytochrome P450 system, is pregnenolone.
2. *3 β -Hydroxysteroid dehydrogenase $\Delta^{4,5}$ isomerase*. This catalyses both the 3 β -hydroxysteroid dehydrogenation and isomerization of the double bond from ring B to ring A. Pregnenolone is thereby converted to progesterone.
- 3/4. *17 α -Hydroxylase/17,20-desmolase*. This catalyses the 17 α -hydroxylation of pregnenolone and progesterone. The desmolase reaction involves the formation of a peroxide at C-20, epoxidation of the C-17 and C-20 carbons and side chain cleavage to form the C-17 oxosteroids dehydroepiandrosterone (from pregnenolone) and androstenedione (from progesterone). Both reactions are catalysed by the same enzyme.



The cyclopentanoperhydrophenanthrene nucleus



Cholesterol

FIGURE 22.4 ■ Structural basis of steroid hormones.

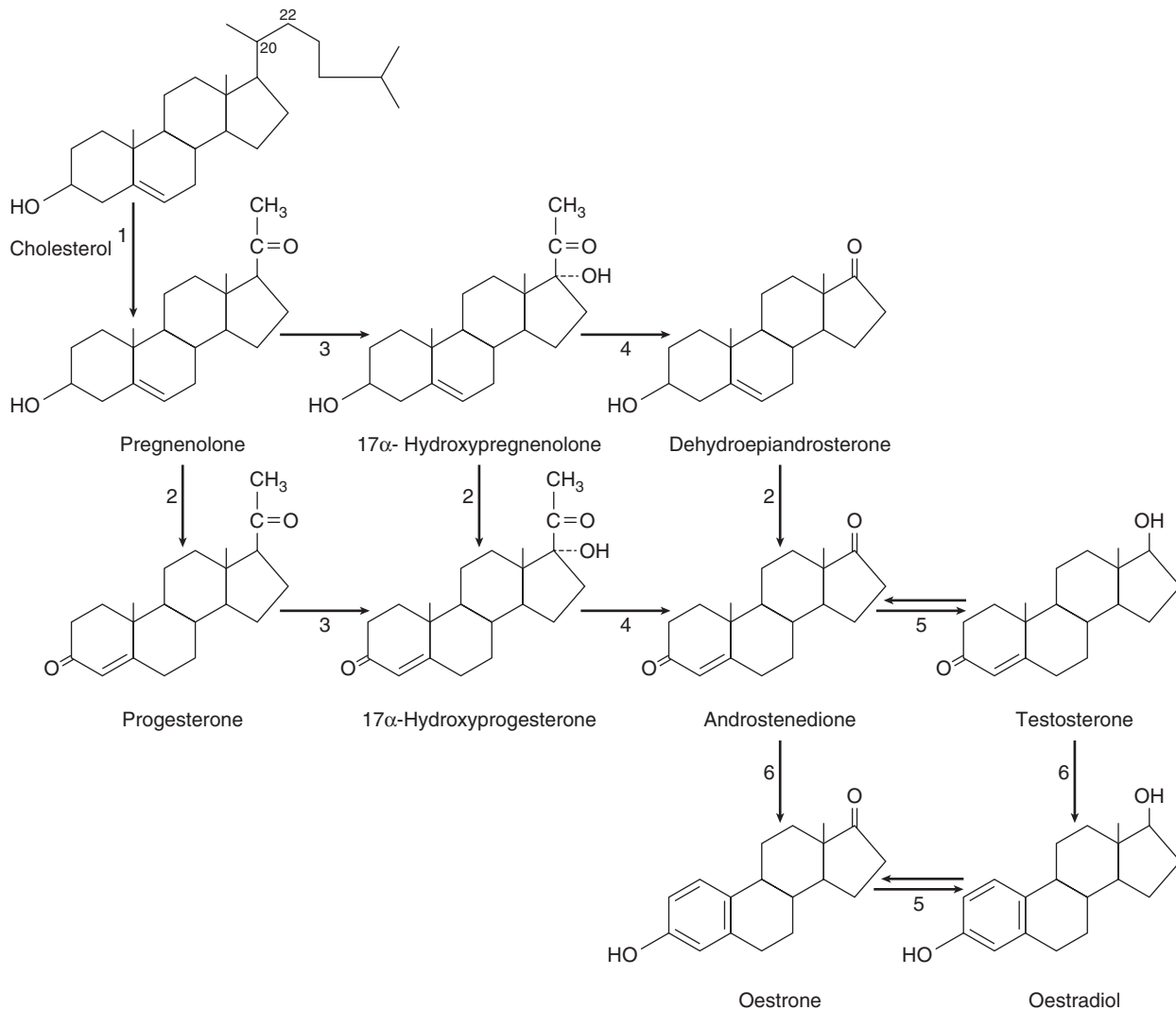


FIGURE 22.5 ■ Ovarian steroid biosynthetic pathways. The numbered arrows correspond to the description of biosynthetic enzymes in the text.

The intermediates are 17 α -hydroxypregnenolone and 17 α -hydroxyprogesterone, respectively.

5. *17-Oxosteroid reductase*. This catalyses the conversion of a 17-oxosteroid to a 17 β -hydroxysteroid and vice versa. Androstenedione and oestrone are converted to testosterone and oestradiol, respectively.
6. *Aromatase*. This converts C-19 $\Delta^{4,3}$ oxosteroids to oestrogens by hydroxylation of the C-19 angular methyl group, oxidation and cleavage of the C-19 methyl group as formaldehyde, dehydrogenation of the A ring, and finally conversion of the 3-oxo group to a 3 β -hydroxy group. The activity of this enzyme in granulosa cells is regulated by FSH.

Oestrogens are synthesized from androstenedione, the major pathway being through oestrone. Pregnenolone is converted to androstenedione either by the Δ^4 pathway (through progesterone) or by the Δ^5 pathway (through dehydroepiandrosterone).

The Δ^4 pathway is favoured in granulosa/corpus luteal cells, while the Δ^5 is favoured in theca cells.

The major secreted oestrogen is 17 β -oestradiol; it is in equilibrium with oestrone in the circulation. Oestrone is further metabolized to oestriol, probably mostly in the

liver. Oestradiol is the most potent oestrogen of the three and oestriol the least.

Steroid secretion through the menstrual cycle

Before ovulation the granulosa cells are not vascularized. The theca cells produce large quantities of androgens, which diffuse into the granulosa cells and are converted to oestrogens. This concept of cooperation between cell types is called the two-cell theory. Vascularization of the granulosa cells leads to increased progesterone secretion in the luteal phase. The stromal cells of the ovaries secrete small amounts of androgens. This secretion becomes more significant after the menopause when it adds to oestrogen concentrations through peripheral conversion in adipose tissue.

Steroid hormone transport and metabolism

Most of the circulating oestrogens and progesterone is protein bound, loosely to albumin and strongly to globulins: oestrogens are carried by SHBG and progesterone

by cortisol-binding globulin. Degradation of both C-18 and C-21 steroids occurs in the liver and involves hydroxylation and water solubilization by conversion to glucuronide and sulphate conjugates. Water-soluble metabolites are excreted in urine and in bile. Quantitative urinary tests were used in the past to assess concentrations of oestrogens and progestogens, but have been superseded by rapid, sensitive and specific serum immunoassays. The major excreted oestrogen and progestogen are oestradiol and pregnanediol, respectively.

Actions of gonadal steroid hormones

Oestrogens

Oestrogens broadly maintain the functions of the reproductive tract, are responsible for the development of secondary sexual characteristics and affect sexual behaviour. Oestrogens cause myometrial hypertrophy, cause the endometrium to thicken in the follicular phase, promote secretion of large amounts of watery cervical mucus around the time of ovulation and maintain vaginal function; after bilateral oophorectomy or the menopause, the uterus and vagina shrink in size.

Oestrogens produce duct growth in the breasts and are responsible for breast enlargement at puberty. Oestrogens, like androgens, increase libido in humans, apparently by a direct action on hypothalamic neurons. In general, oestrogens reduce secretion of LH and FSH (negative feedback) but, in mid-cycle, oestrogens increase LH secretion (positive feedback). In addition to their reproductive role, oestrogens have important systemic effects: they maintain bone density and skin thickness and protect against atheromatous arterial disease in premenopausal females.

Synthetic oestrogen derivatives with agonist actions are used in contraceptive and hormone replacement preparations. Ethinylloestradiol is the most commonly used, particularly in contraceptives; it is potent and, unlike naturally occurring oestrogens, is active when given by mouth.

Anti-oestrogen preparations are also available: clomifene citrate and tamoxifen are non-steroidal triphenylethylene derivatives with mixed oestrogen agonist and antagonist actions. Clomifene citrate is used to induce ovulation, and tamoxifen to treat oestrogen-dependent breast cancers.

Progestogens

Progesterone causes secretory changes in endometrium already primed by oestrogen. It promotes the secretion of smaller amounts of thicker cervical mucus. It is essential for the maintenance of early pregnancy.

Synthetic steroids with progesterone agonist properties are used in contraception and hormone replacement therapy. Those commonly used tend to be derivatives either of 17 α -hydroxyprogesterone or of 19-nortestosterone. Both types of derivative are used in combined oral contraceptive preparations: along with ethinylloestradiol, they prevent follicle growth, promote an endometrial reaction unfavourable to implantation and render cervical mucus thick and impenetrable to sperm.

Mifepristone is a derivative of norethisterone that blocks the actions of progesterone peripherally. It induces

menstruation when given in the late luteal phase and induces abortion when given in pregnancy.

Androgens

The two main androgens in the female are testosterone and androstenedione. Dehydroepiandrosterone (DHEA) and its sulphate (DHEAS) are less important androgens. During the reproductive years, 90% of DHEA and DHEAS is synthesized by the adrenals and 10% by the ovaries. Androstenedione is derived equally from the ovaries and adrenals. Only 50% of testosterone production is glandular (roughly equal contributions from adrenals and ovaries), the remainder resulting from peripheral metabolism of weaker androgens, mostly in adipose tissue. A total of 99% of testosterone is bound and 1% is free: the great majority (78%) is bound to SHBG. This is a glycoprotein synthesized in the liver, with a carbohydrate content of 34%. It is a high-affinity, low-capacity binding protein. Plasma SHBG concentrations are raised when oestrogen concentrations are high (as in pregnancy) and in hyperthyroidism; concentrations are low in women treated with testosterone. Because SHBG-bound testosterone is relatively inert biologically, it may be useful for laboratories to measure both testosterone and SHBG, to give an indication of how much testosterone is not bound to SHBG ('free androgen index').

OLIGO- AND AMENORRHOEA

Oligomenorrhoea is defined as a menstrual cycle length of greater than six weeks but less than six months. Amenorrhoea is complete absence of menstruation or cycle length greater than six months. In these conditions, ovulation does not occur or is very infrequent. Women with oligo- or amenorrhoea may seek medical assistance because their bleeding pattern is abnormal, because of infertility, because of hirsutism/virilism or with a combination of these complaints.

The majority of patients with oligomenorrhoea have the polycystic ovary syndrome. The classic 'polycystic ovary syndrome' (PCOS) was described by Stein and Leventhal in 1935; they described an association between polycystic ovaries and oligo- and amenorrhoea in a small series of patients, most of whom were infertile or hirsute. It later became apparent that plasma concentrations of LH and/or androgens are elevated in many, but not all, women with polycystic ovaries; FSH concentrations are normal. However, it should be noted that around a quarter of women who have no gynaecological symptoms have also been found to have polycystic ovaries.

Polycystic ovary syndrome is a syndrome of ovarian dysfunction. It seems most likely that changes in the pattern of gonadotrophin secretion are responsible for the failure of antral follicles to grow and for a dominant follicle to emerge and ovulate (Fig. 22.2A). It is possible to induce follicular growth in patients with PCOS by therapeutic administration of FSH. It is important to recognize that polycystic ovaries and PCOS can occasionally occur as a secondary phenomenon in patients with other endocrine disorders. These include conditions such as

congenital adrenal hyperplasia and Cushing syndrome, hyperprolactinaemia and acromegaly.

The revised diagnostic criteria of PCOS require any two of the following to be present:

- oligo- or anovulation
- clinical and/or biochemical evidence of hyperandrogenism
- polycystic ovaries

and exclusion of other aetiologies (e.g. congenital adrenal hyperplasia, androgen-secreting tumours, Cushing syndrome).

The primary clinical indicator of androgen excess is the presence of hirsutism or acne. The sole presence of androgenic alopecia as an indicator of hyperandrogenism has been less well studied.

Few data are available on the value of routinely measuring androstenedione in hyperandrogenic patients, although it may be somewhat more elevated in patients with non-classic 21-hydroxylase deficiency congenital adrenal hyperplasia than in PCOS. Nonetheless, at present, the paucity of clinical data with androstenedione precludes its recommendation for the routine assessment of hyperandrogenism.

Weight loss can help to regularize menstrual periods with resumption of natural ovulation. If weight loss fails, clomifene, tamoxifen or gonadotrophins may be used to induce ovulation. Metformin may also improve ovulation rates when used with clomifene in patients resistant to clomifene alone.

Cyproterone acetate (an anti-androgenic progestogen) reduces hirsutism and is usually given as part of a combined oral contraceptive pill. The oestrogen component increases the plasma concentration of SHBG, which is often low in women with PCOS. The result is a decrease in the concentration of free testosterone, which also contributes to the reduction in hirsutism and acne. The treatment adopted depends on the desired outcome, for example whether this is to control acne, regularize menstruation or to achieve pregnancy.

Polycystic ovary syndrome is frequently associated with the metabolic syndrome of insulin resistance, hyperinsulinaemia and dyslipidaemia. This topic is discussed in detail in Chapter 15.

Hyperprolactinaemia is another important cause of oligo/amenorrhoea and acts by suppressing LH secretion and inhibiting ovulation. The normal range for plasma prolactin concentration in women is approximately 300–700 mIU/L, although this is dependent on the assay used. It is difficult to define a precise upper limit, but significant disease is unlikely unless the prolactin concentration is >1000 mIU/L. The possibility that hyperprolactinaemia is secondary to drugs (e.g. opiates, methyl dopa, antipsychotics, oestrogens and metoclopramide) must be excluded. Thyrotropin-releasing hormone (TRH) stimulates prolactin secretion, and therefore primary hypothyroidism can cause oligomenorrhoea. Patients found to have unexplained significant hyperprolactinaemia should have pituitary imaging to exclude a macroprolactinoma (a prolactin-secreting adenoma) or a non-functioning tumour (interfering with the normal dopamine suppression of prolactin secretion); the most frequent diagnoses resulting are idiopathic

disease (no scan abnormality) or a microadenoma. (See Chapter 18 for further details.)

The medical management of hyperprolactinaemia is with a dopamine agonist. Bromocriptine has been used widely, but cabergoline, a longer acting preparation, is now usually preferred. It is better tolerated and more efficacious than bromocriptine but is associated with psychiatric adverse effects. Quinagolide is another longer acting preparation, and an alternative to cabergoline. This topic is discussed further in Chapter 18. Trans-sphenoidal surgery is an option for some patients.

Primary ovarian failure also causes amenorrhoea. It is diagnosed by the finding of a raised plasma FSH concentration (>40 IU/L) and is irreversible. Other causes of oligomenorrhoea include androgen-secreting tumours and late onset congenital adrenal hyperplasia: these conditions are discussed later in this chapter, in the account of hirsutism and virilism.

INFERTILITY

Infertility is defined as failure to conceive after 12 months of regular sexual intercourse without contraception. The partner's semen should be analysed to exclude male subfertility, but the results are often inconclusive; insemination of mature oocytes in vitro establishes the fertilizing capacity of semen more reliably.

Imaging techniques are of considerable importance in the investigation of infertility. Laparoscopy with hydrotubation, hysterosalpingography and hysterocontrast sonography all assess patency of the fallopian tubes, using direct vision, radiography and sonography respectively; intraperitoneal spillage of fluid injected through the cervix excludes tubal blockage. Transvaginal ultrasonography is used to assess ovarian morphology and to monitor ovulation.

Tests to predict ovulation are required in women who have long cycles (>35 days). The LH surge that precedes follicle rupture can be detected by qualitative urine tests based on monoclonal antibodies. Commercially available point-of-care tests are very sensitive and are increasingly being used. Measuring serum progesterone concentration in the mid-luteal phase (i.e. seven days after time of expected ovulation) is a well-established method for assessing ovulation. Values in conception cycles have been found to range between 28 and 53 nmol/L; 30 nmol/L is generally taken as the lower limit of normal for mid-luteal phase progesterone. Blood sampling before or after the mid-luteal peak can give misleadingly low values. Sampling can be timed correctly by taking serial specimens and identifying the mid-luteal phase specimen retrospectively. It is the specimen taken seven days before the first day of the next menstrual period. (The classic day 21 sample will only be the mid-luteal phase if the woman has a regular 28-day cycle.) Progesterone is secreted in a pulsatile fashion during the luteal phase, concentrations varying by as much as a factor of three; when sampling coincides with a trough, the plasma progesterone concentration may appear abnormally low.

Basal plasma concentrations of FSH and LH are measured routinely in infertile women. A single specimen is

usually taken during the early follicular phase on day 1–3 of the cycle. Interpretation of apparently abnormal early follicular phase gonadotrophin concentrations must take into account the pulsatile nature of release of both hormones. For LH, the pulse interval in the early follicular phase is 65–90 min, with peak and trough concentrations varying by as much as a factor of five. Plasma FSH concentrations do not fluctuate so widely – by a factor of two at the most. Elevated FSH concentrations in women with normal-length menstrual cycles indicate diminished fertility and possibly incipient ovarian failure. In patients undergoing assisted conception, early follicular phase plasma FSH concentrations are inversely related to the number of oocytes obtained and to the probability of conception. Measurement of AMH, however, is now preferred, as it gives a more accurate reflection of ovarian reserve and likely response to ovarian stimulation, and is not dependent on the day of the cycle. Measurement of early follicular phase LH is of less certain diagnostic value; concentrations tend to be abnormally high in patients with PCOS. Misleadingly high values are obtained when sampling coincides with a pulse peak. Basal concentrations of prolactin and thyroid hormones are unlikely to be of diagnostic value unless the length of the menstrual cycle is abnormal.

Increasingly, subfertility, whatever its cause, is treated using assisted conception techniques. Modern superovulation regimens promote the development of more than one follicle and include the use of gonadotrophin releasing hormone analogues, follicle stimulating hormone and human chorionic gonadotrophins.

Oocytes are collected by ultrasound-directed follicle aspiration. They are then either incubated with sperm (in vitro fertilization, IVF) or directly injected (intracytoplasmic sperm injection, ICSI). The resulting embryos are then transferred to the uterine cavity.

The major risk of superovulation regimens is ovarian hyperstimulation syndrome (OHSS). This generally occurs in the luteal phase or during early pregnancy and is triggered by either exogenous or endogenous hCG. Patients with polycystic ovaries, more than 20–25 follicles in both ovaries or with a high (>6000 pmol/L) or rapidly rising oestradiol (doubling on consecutive days) prior to hCG administration are most at risk. Mild OHSS is common (20–30% of stimulated cycles) and leads to abdominal distension, enlarged ovaries, nausea and vomiting and is treated conservatively. Severe OHSS (<2%), caused by increased vascular permeability and fluid shift out of the intravascular space, presents with ascites, increased blood viscosity, oliguria, thromboembolism and respiratory problems and can result in multi-organ failure. Biochemical changes include hyponatraemia from hypersecretion of antidiuretic hormone, hyperkalaemia, hypoalbuminaemia, abnormal liver enzymes and high plasma creatinine.

HIRSUTISM AND VIRILISM

Hirsutism means excessive hair growth in sites usually associated with male sexual maturity, that is, the face, lower abdomen (giving the appearance of a male escutcheon),

anterior thigh, mid-chest and periareolar region. It can be associated with acne and oily skin. Virilism refers to more extreme manifestations of androgen exposure, including temporal hair recession, clitoromegaly, increased muscle mass, breast atrophy and deepening of the voice.

Dihydrotestosterone derived from testosterone acts on hair follicles to stimulate growth of terminal hair (long, coarse and pigmented) rather than vellus hair (short, fine and poorly pigmented). Once stimulated, the hair follicle remains responsive to low-dose weak androgen exposure because it has attained the ability to convert weak androgens to testosterone.

Polycystic ovary syndrome is responsible for hirsutism in a substantial proportion of patients. Rarer causes are discussed below. In many cases, however, no cause can be diagnosed; such hirsutism is termed idiopathic.

In idiopathic hirsutism, total plasma testosterone concentrations are normal, but subtle androgen disorders can generally be identified. Plasma androstenedione concentrations may be elevated and SHBG reduced; free testosterone concentration may be raised even though total testosterone is within normal limits.

The investigation of hirsutism should determine whether androgen concentrations are indeed elevated and to what degree, and rule out serious disease (particularly malignancy) of the ovaries or adrenals. Grossly elevated testosterone concentrations (>5 nmol/L) and hirsutism of sudden onset (especially if accompanied by virilism) suggest malignancy.

Computed axial tomography (CT) or magnetic resonance imaging (MRI) is used to visualize adrenal tumours. Laboratory investigations used to evaluate hirsutism include measurements of serum cortisol and 17 α -hydroxyprogesterone, in addition to testosterone, androstenedione and SHBG; steroid hormone concentrations following overnight dexamethasone suppression or ACTH stimulation may also be useful. Congenital adrenal hyperplasia is a recognized cause of hirsutism. The most frequent type is 21-hydroxylase deficiency, which results in elevated plasma concentrations of 17 α -hydroxyprogesterone and increased synthesis of adrenal androgens. Testosterone concentrations are significantly elevated and ACTH stimulation is followed by an excessive increase in 17 α -hydroxyprogesterone (see [Appendix 22.1](#), below, for protocol). In Cushing syndrome, there is typically loss of the normal diurnal rhythm in plasma cortisol concentrations and failure of suppression in the overnight dexamethasone suppression test.

Although 90% of DHEAS is normally of adrenal origin, elevated DHEAS concentrations in hirsutism do not confirm an adrenal source of the excess: elevated DHEAS concentrations may also be found in polycystic ovary syndrome. However, a DHEAS concentration of more than twice the upper limit of normal does suggest the possibility of an adrenal tumour. Suppression of elevated androgen concentrations in hirsutism by dexamethasone might be expected to confirm an adrenal source; however, selective venous catheterization studies have demonstrated that elevated ovarian androgens can also be suppressed by dexamethasone in patients with hirsutism.

Once the rarer causes of hirsutism ([Box 22.1](#)) have been excluded, treatment consists of medication to suppress

BOX 22.1 Some rare causes of hirsutism

- Congenital adrenal hyperplasia (adult onset)
- Cushing syndrome
- Ovarian tumours
- Adrenal tumours
- Self-administration of androgens

androgen secretion or counter its effects. Combined oral contraceptives and the anti-androgen cyproterone acetate are widely used. Medical treatment can be supplemented by cosmetic techniques such as shaving, waxing and electrolysis.

PREGNANCY**Introduction**

Pregnancy is associated with marked hormonal changes in the maternal circulation that facilitate the metabolic, vascular and immunological adjustments necessary for the fetus to thrive. Abnormal concentrations of these hormones, or of other plasma constituents of fetoplacental origin, may indicate gestational pathology. The metabolic changes associated with pregnancy are relatively short lived and thus are rarely harmful to the healthy mother.

Prior to the widespread availability of high-resolution ultrasound scanning, the diagnosis of early pregnancy failure (such as spontaneous abortion and ectopic pregnancy), fetal malformation and fetal growth disturbances relied solely on clinical evaluation backed by tests on maternal blood. Although these indirect and often inaccurate methods have largely been replaced by ultrasound technology, biochemical assays still form an important part of the screening of many pregnancies.

Biochemical diagnosis of pregnancy**Human chorionic gonadotrophin**

Human chorionic gonadotrophin (hCG) can be detected in maternal blood 7–9 days after conception (i.e. 22–24 days after the last menstrual period in women with regular 28-day cycles) and in urine 1–2 days later. The biochemical diagnosis of pregnancy is routinely made by a monoclonal antibody enzyme immunoassay that detects urinary hCG concentrations as low as 25 IU/L. The plasma concentration of hCG in normal pregnancy rises steeply to a peak at ten weeks of gestation and then falls and remains at a lower concentration for the remainder of pregnancy. This fall at the end of the first trimester is peculiar to hCG, as concentrations of the other chemical products of the fetoplacental unit rise with increasing length of gestation.

High sensitivity assays have shown that hCG is present in the plasma at low concentrations (<5 IU/L) prior to conception. The biochemical diagnosis of pregnancy should only be made when plasma hCG concentration exceeds 25 IU/L or if a lower concentration doubles within two days. In cases where exogenous hCG has been given

to induce ovulation, hCG estimations should be delayed by up to 14 days to allow clearance of the administered hCG.

The use of quantitative serum hCG estimations in the diagnosis of ectopic pregnancy is discussed below.

Apart from its place in the diagnosis of pregnancy, other well-established roles for hCG estimation are in the diagnosis and monitoring of gestational trophoblastic disease and as a marker for some tumours (see Chapter 42). The diagnosis of gestational trophoblastic disease is usually made ultrasonographically or histologically after an episode of vaginal bleeding in early pregnancy, though the condition may present after a term pregnancy when persistently high concentrations of hCG may help in making the diagnosis. After the uterus has been surgically evacuated, persistence or recurrence of the condition is monitored by serial quantitative hCG assays. The response of gestational trophoblastic disease to chemotherapy can be monitored in a similar fashion.

Human chorionic gonadotrophin measurement also plays a part in screening for Down syndrome. This topic is discussed on p. 443.

Diagnosis of ectopic pregnancy

Ectopic pregnancy poses a considerable hazard to the mother (95% are tubal: a ruptured tubal pregnancy can cause severe intra-abdominal bleeding), so that accurate diagnosis of this condition is essential. The symptoms and signs are notoriously variable and definitive diagnosis frequently depends on laparoscopy. However, use of additional investigations, such as measurement of hCG concentration, may help to reduce the number of ‘unnecessary’ laparoscopies and allow medical or conservative treatment.

Transvaginal ultrasound enables the diagnosis of intrauterine pregnancy at a time when the plasma hCG concentration is around 1000 IU/L. The absence of a visible embryonic sac within the uterus in these circumstances makes an ectopic pregnancy the likely diagnosis. The rate of increase in plasma hCG concentrations may provide useful additional information. Early in a normal pregnancy, these usually double every two days, while the rise in ectopic pregnancy is, in general, less. Progesterone measurements have also been used in the investigation of possible ectopic pregnancy. A plasma progesterone concentration of >25 nmol/L is suggestive of a viable intrauterine pregnancy; however, recent guidelines from the National Institute for Health and Care Excellence (NICE) in the UK suggest that progesterone measurement should no longer be used in this context. Low or falling concentrations of progesterone and hCG may allow expectant management if symptoms are minimal, as this suggests that the pregnancy is failing and will resolve spontaneously. Surgery can also be avoided in patients with hCG <5000 IU/L and minimal symptoms, as they can be treated with the folate antagonist, methotrexate, as long as hCG monitoring is available. Pain and haemodynamic instability are indications for laparoscopy.

In clinical practice, the diagnosis of ectopic pregnancy depends on a high index of suspicion, measurement of serum hCG concentration, pelvic ultrasound and, when indicated, laparoscopy.

Biochemical monitoring of pregnancy

Spontaneous abortion

Many studies have reported low concentrations of various hormones and placental proteins in patients with early spontaneous abortion. However, these studies have often been performed after the onset of bleeding and their findings may have been the result, rather than the cause, of the pregnancy failure. Plasma progesterone concentrations have, for example, been reported to be low in these circumstances, but administration of progestogens to women with threatened abortions has not been shown to prevent pregnancy loss.

Screening for fetal malformation

Prior to the widespread availability of high-resolution ultrasound scanners, the antenatal detection of fetal malformation was based on maternal biochemical screening. To some extent, ultrasound has made this screening less important, but these tests may focus attention on the high-risk pregnancy so that an appropriate diagnostic procedure can be performed.

Biochemical screening. Screening for Down syndrome (trisomy 21) in the past has depended entirely on maternal age and past history of chromosomal abnormalities in offspring and has resulted in some form of invasive tissue sampling (chorionic villus sampling or amniocentesis) being offered to essentially all pregnant women over a locally agreed age, usually 35 years. (The risk of a Down syndrome pregnancy is 1:1528 at maternal age 20, 1:1351 at 25, 1:909 at 30, 1:384 at 35 and 1:112 at age 40.) This approach was unsatisfactory because, although the risk of trisomy 21 increases with maternal age, 70% of Down infants are born to women below the age of 35. In addition, many women over 35 decline the offer of invasive tissue sampling because of the risk of miscarriage.

The observation in 1984 that maternal plasma α -fetoprotein (AFP) concentrations were lower in the presence of fetal Down syndrome than in unaffected pregnancies, led to its use in screening for this condition. Subsequently, maternal plasma concentrations of hCG were found to be elevated and concentrations of unconjugated oestriol decreased when the fetus had Down syndrome. The assay of these substances in maternal plasma at around 16 weeks of gestation has become known as the 'triple test' and, in combination with maternal age and weight, gives a risk score of the fetus having Down syndrome.

The chosen cut-off for a 'positive' screen will govern the detection and false positive rates. A commonly chosen cut-off is a triple test result representing a risk of 1 in 250 or higher, corresponding approximately with the risk for a maternal age of 35 years alone. At this cut-off, there is a detection rate of 61% and a false positive rate of 5%.

Maternal biochemical screening for Down syndrome is not diagnostic of the condition, but allows a better selection of a high-risk group (to whom diagnostic amniocentesis may be offered) than use of maternal age alone. However, screening does not detect other, more serious trisomies, for which older women are also at increased

risk. The detection and analysis of free fetal DNA in the maternal plasma may, in future, obviate the need for invasive testing and reduce the risk of fetal loss.

α -Fetoprotein is one of the two major proteins in the fetal circulation. The AFP concentration in fetal plasma greatly exceeds that in amniotic fluid and maternal plasma. Thus, even minor leakages from the conceptus result in readily detectable rises in maternal plasma AFP concentrations. This formed the basis of maternal screening in the second trimester for a range of fetal malformations, e.g. neural tube defects, but these are now reliably diagnosed by ultrasound. Several situations can give rise to false positive results, the most important of which is an incorrect estimation of gestational age, as maternal plasma AFP rises steeply at 16–20 weeks (Fig. 22.6). Multiple pregnancies are also associated with higher concentrations.

Ultrasound. While ultrasound examination allows the diagnosis of the majority of significant structural fetal anomalies, most genetic disorders cannot be diagnosed by this means and require tissue sampling techniques to identify an affected fetus. The majority of these procedures are performed for fetal karyotyping, but despite the increasing use of tissue sampling techniques such as amniocentesis and chorionic villus sampling (CVS) over the last 20–30 years, there has not been a significant reduction in the birth incidence of infants with abnormal karyotypes.

Current screening practice. An important advance in the antenatal diagnosis of Down syndrome was the recognition of the nuchal translucency (NT) sign. This is the appearance on ultrasound of increased fluid at the back of the fetal neck in affected fetuses and can be detected late in the first trimester. Its specificity and sensitivity are enhanced if it is combined with other investigations such as measurement of hCG and pregnancy-associated plasma protein A (PAPP-A; the 'combined test'). Pregnancy-associated plasma protein A is a zinc metalloproteinase that acts on insulin-like growth factor binding protein 4, decreasing its affinity for IGF-1 and IGF-2, and thus regulating IGF activity in certain tissues. The combined test

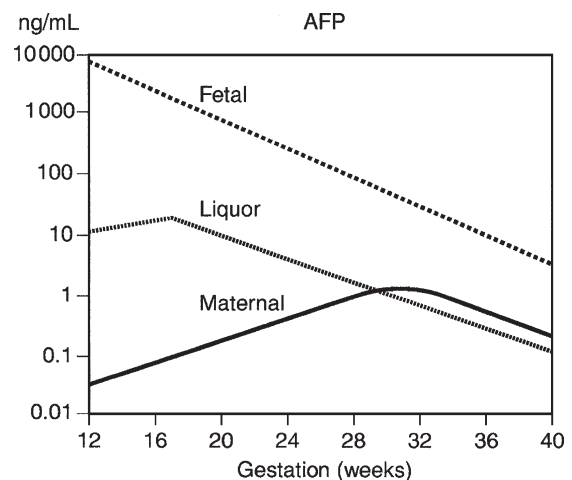


FIGURE 22.6 ■ Changes in α -fetoprotein (AFP) concentration with gestation in maternal and fetal plasma and amniotic fluid liquor.

BOX 22.2 Screening for Down syndrome

- From 11 to 14 weeks:
 - Nuchal translucency (NT)
 - The combined test (NT, hCG and PAPP-A)
- From 14 to 20 weeks:
 - The triple test (hCG, AFP, OE3 (oestriol))
 - The quadruple test (HCG, AFP, OE3 and inhibin A)
- From 11 to 14 *and* 14 to 20 weeks:
 - The integrated test (NT, PAPP-A+hCG, AFP, OE3, inhibin A)
 - The serum integrated test (PAPP-A+hCG, AFP, OE3, inhibin A)

Abbreviations are explained in the text.

detects 90% of affected pregnancies with a screen positive rate of 2%, if a cut-off risk of 1:150 is chosen.

For women presenting only in the second trimester, the recommended screening tests are either the 'triple test', involving measurement of hCG, AFP and oestriol or the 'quadruple test' – the triple test with the addition of inhibin A. Other available tests are: the 'integrated test' (NT and serum measurements) or, if NT is not available, measurements on maternal serum alone ('the serum integrated test') (see [Box 22.2](#)). The integrated test (if good quality NT measurement is available) performs better than the combined test but is more difficult to administer. The serum integrated test performs less well.

Women undergoing screening should appreciate that positive tests require confirmation (diagnosis) by an invasive procedure: before 12 weeks of gestation, this is by chorionic villus sampling, thereafter by amniocentesis.

Fetal tissue sampling techniques

The prenatal diagnosis of many chromosomal and genetic abnormalities requires tissue sampling. Various techniques are available (see [Table 22.1](#) and [Fig. 22.7](#)) but, with few exceptions, if a diagnostic test is available for the condition under investigation, any of the methods of sampling may be employed. The differences between the techniques are primarily the gestation at which the test may be performed and the delay between sampling and the result. The list of inherited disorders that can be diagnosed by DNA studies is growing rapidly. However,

TABLE 22.1 Techniques for assessment of the fetus

Procedure	Gestation	Delay in result	
		PCR	Culture
Preimplantation diagnosis	3 day embryo	See text	–
Chorionic villus sampling	10+ weeks	2 days	2 weeks
Amniocentesis	15+ weeks	3 days	3 weeks
Cordocentesis	19+ weeks	–	1 week
Fetoscopy (anatomy/sample)	15–20 weeks	–	1 week

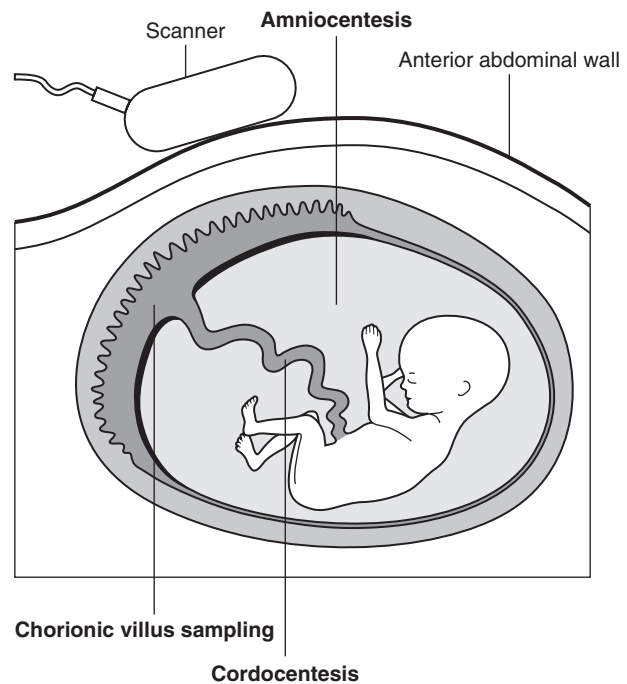


FIGURE 22.7 ■ Ultrasound-guided tissue sampling sites used during pregnancy.

in many cases, these conditions are not due to single gene abnormalities, or disease-specific gene probes are not available, and family studies may be necessary to evaluate individual cases to assess their suitability for prenatal diagnosis. In vitro fertilization has also opened up the possibility of preimplantation genetic diagnosis from one or two cell biopsy of the embryo, using comparative genomic hybridization or single nucleotide polymorphism arrays. It is beyond the scope of this book to discuss fetal tissue sampling techniques in detail, and this section only summarizes the essential features.

Chorionic villus sampling

As the fetus and the placenta both develop from the same early blastocyst, their genetic make-up is identical in the vast majority of cases. Thus, chromosomal and DNA analysis of the placenta (chorionic villi) will provide information about the fetus. The post-procedure spontaneous miscarriage rates for transabdominal and transcervical CVS are comparable at approximately 2.5%, not significantly different from the background rate of loss. Potential sources of diagnostic inaccuracy from CVS are maternal cell contamination and placental mosaicism.

Amniocentesis

Traditionally, amniocentesis was usually performed at 15–16 weeks of gestation, but it can be performed earlier. Results of cell culture become available 2–4 weeks later (though a preliminary result may be available within hours from direct examination); the associated fetal loss rate is approximately 1%. Amniocentesis can be used for biochemical assessment of the liquor for fetal lung maturity (now rarely used in the UK). The production of the

surfactants lecithin and sphingomyelin by the fetal lung are almost equal until ~32 weeks of gestation, after which the lecithin concentration increases more rapidly than that of sphingomyelin. A lecithin–sphingomyelin ratio of >2 indicates a low risk of the baby developing respiratory distress syndrome after birth. Amniocentesis can also be used for the monitoring of haemolytic disease of the fetus by determining the Rhesus genotype on uncultured amniocytes and measuring the optical density of the liquor at 450 nm as an indication of bilirubin concentration.

Cordocentesis

Fetal blood can be obtained from 19 weeks of gestation by cordocentesis, an outpatient procedure that does not require maternal sedation or the use of fetal paralysis. After visualization of the insertion of the umbilical cord into the placenta by real-time ultrasound, a needle is passed into the umbilical cord through the (locally anaesthetized) maternal abdomen. The fetal loss rate after cordocentesis is approximately 1%.

Monitoring of maternal and fetal well-being

The increased accuracy and availability of ultrasound-based biophysical tests of fetal well-being has resulted in a marked decline in the use of biochemical tests, except for a few clinical situations, for example serum uric acid measurements in a patient with pre-eclampsia, and 'liver function' tests and serum bile acid measurements in women with obstetric cholestasis.

Pre-eclampsia is the occurrence of proteinuria and hypertension in pregnancy; peripheral oedema is often present. Untreated, it may progress to life-threatening severe hypertension, renal failure and convulsions (eclampsia) or be complicated by the haemolysis, elevated liver enzymes and low platelets (HELLP) syndrome. An early biochemical feature of pre-eclampsia is an increase in plasma urate concentrations, as a result of increased renal tubular reabsorption of urate secondary to decreased renal perfusion. However, this is an unreliable indicator, as urate concentrations rise in the last trimester of normal pregnancies.

Cholestasis of pregnancy is discussed in more detail in Chapter 14. The term includes acute fatty liver of pregnancy (of which hyperuricaemia is also a feature) and HELLP syndrome.

Intrapartum fetal monitoring

The widespread use of fetal heart rate (FHR) monitoring in labour resulted in an increase in the number of operative deliveries, frequently delivering a neonate without any sign of asphyxia. As a result, fetal scalp blood sampling was introduced in order to improve the specificity of FHR monitoring alone. A variety of studies have demonstrated that, when interpreting FHR patterns, variability and accelerations are the features that correlate most closely with fetal scalp pH or lactate concentration. The presence of normal variability or accelerations is indicative of normal oxygenation, whereas the presence of reduced variability or decelerations makes poor oxygenation and

hence acidosis, more likely. Although there is good evidence that continuous FHR monitoring can detect a low fetal pH, acidosis, especially in well-grown term fetuses, does not correlate well with fetal death or cerebral injury. Indeed, a number of studies suggest that only 8–15% of cases of cerebral palsy are associated with events occurring during labour and, even in these, the link is not proven as causal.

Biochemical changes during pregnancy

Pregnancy is a major endocrine event in the female lifespan, involving wide-ranging and often considerable changes in the metabolism of various hormones. Concentrations of oestrogens, progesterone, testosterone and prolactin all increase steadily through gestation, while those of LH and FSH decrease. Sharp increases in the concentrations of hormone-binding globulins such as SHBG are seen, reaching a plateau by the end of the first trimester. In addition, placental hormones such as hCG and human placental lactogen (hPL) appear in the maternal circulation within the first few weeks following conception. Pregnancy does not usually alter thyroid function to a significant extent, but changes in individual indices of thyroid function are often seen, such as increases in plasma concentrations of total thyroxine (T4) secondary to an increase in thyroxine binding globulin. Free T4 and free tri-iodothyronine (fT3) concentrations may rise slightly in early pregnancy (as a result of the thyrotrophic action of hCG), but later fall to within, or just below, the non-pregnant reference ranges. Thyroid stimulating hormone concentrations tend to fall in early pregnancy (sometimes to undetectable concentrations), but later return to normal.

Associated with the changes in gonadotrophins and steroid hormones, pregnancy affects several metabolic processes in various tissues and organs, in turn altering the concentrations of many maternal plasma constituents. In some instances, these changes are apparently unrelated to the needs of mother or fetus. Because the use of standard reference ranges is often inappropriate during pregnancy, an appreciation of the metabolic effects of pregnancy is desirable.

An important consideration when interpreting the effects of pregnancy on blood components is the associated change in plasma volume. Total body water increases by about 7 L during pregnancy, with an increase in extracellular fluid being responsible for at least half of this increase. Marked alterations in renal haemodynamics accompany this increase in extracellular fluid. Both glomerular filtration rate (GFR) and renal plasma flow increase by up to 50% above pre-pregnancy values, changes that are compensated for by changes in tubular reabsorption. Plasma creatinine and urea concentrations decrease during pregnancy, and increases in creatinine clearance (paralleling the increase in GFR) become apparent four weeks after conception, peaking 9–11 weeks later. Renal function may deteriorate markedly during pregnancy in women with pre-existing renal disease or a history of pre-eclampsia in a previous pregnancy, and monitoring of plasma urate and creatinine concentrations is advisable.

Plasma proteins

Pregnancy induces widespread changes in plasma protein concentrations, owing both to the effect of changes in plasma volume and to specific hormone-induced changes in protein synthesis and degradation. In general, total protein concentrations are slightly reduced, with those of albumin falling by approximately 15% and those of globulins (especially acute phase proteins) often increasing. Two to three-fold increases in the concentrations of α_1 -antitrypsin, fibrinogen and caeruloplasmin are often seen during the later stages of gestation. Prealbumin is a notable exception as plasma concentrations fall during pregnancy. Plasma immunoglobulin concentrations are rarely affected by pregnancy.

Pregnancy may induce slight increases in the plasma activity of aminotransferases such as aspartate aminotransferase. Substantial increases in alkaline phosphatase activities occur, reaching a peak of 2–3 times normal adult values during the third trimester. This increase is largely due to the presence of the placental isoenzyme in maternal plasma. Although changes in the fetal isoenzyme have been related to problems of fetal development, this has proved of little diagnostic use owing to the wide range of isoenzyme activities seen in normal pregnancy.

Plasma lipids and lipoproteins

The hyperlipidaemia of pregnancy is well recognized. By the third trimester, fasting plasma triglyceride concentrations are increased 2–3 times, with lesser increases being seen in total cholesterol, phospholipids and non-esterified fatty acids. These increases result from a complicated sequence of changes in the major plasma lipoprotein classes, low density lipoproteins (LDL) and high density lipoproteins (HDL). LDL concentrations rise steadily, but level off during the last ten weeks of pregnancy, whereas those of HDL initially increase, but level off (and may even fall) approximately half way through pregnancy; this may be a result of the onset of the insulin resistance that is typical of pregnancy. Serial measurement of hormones during gestation indicates that the hyperlipidaemia of pregnancy is influenced not only by oestradiol and progesterone, but also by insulin and hPL.

Glucose tolerance

Fasting plasma glucose concentrations usually fall slightly during pregnancy, whereas postprandial concentrations increase. Pregnancy may lead to a deterioration in glucose tolerance, which is subtle in the majority of women, probably due to insulin resistance induced by elevated concentrations of sex hormones. In patients with pre-existing type 1 diabetes mellitus, this may result in an increased requirement for exogenous insulin. The insulin requirement should thus be monitored closely (by regular measurements of blood glucose and glycated haemoglobin) in order to reduce the incidence of fetal malformation, macrosomia and intrauterine death, conditions that are all associated with poor glycaemic control.

The precise nature of the defect in gestational glucose intolerance remains controversial. Patients who are initially free from diabetes but develop gestational diabetes may

either revert to normal glucose tolerance after pregnancy (in which case they remain at risk of developing diabetes in the future) or remain frankly diabetic. Although screening for gestational diabetes has been advocated (based on factors such as obesity, strong family history of diabetes mellitus or previous delivery of a baby weighing 4 kg or more), most 'at-risk' patients do not develop glucose intolerance and those that do often have low risk factor scores. The presence of glycosuria is similarly of little use, as the majority of pregnant women will have glycosuria at some stage of their pregnancy owing to their lowered renal threshold, rather than glucose intolerance per se.

Consequently, routine measurement of blood glucose has been advocated in *all* pregnancies. While this will indeed detect patients with gestational glucose intolerance, controversy exists concerning the indications for subsequent formal glucose tolerance testing, as well as the stage of gestation at which such testing should take place. In addition, there is no agreement as to what constitutes an abnormal tolerance test profile, as the World Health Organization (WHO) criteria are based on non-pregnant individuals and other criteria are purely statistical (e.g. values more than two standard deviations above the mean), rather than relating to the outcome of pregnancy. In the UK, the National Institute for Health and Care Excellence (NICE) has produced guidelines, including for the selection of patients for formal glucose tolerance testing. Gestational diabetes, and the management of maternal diabetes during pregnancy are discussed in Chapters 15 and 16.

Other changes

Mild proteinuria may accompany the changes in plasma protein concentrations seen during pregnancy. However, this may be an early warning of a threat to mother and fetus if it develops into pre-eclampsia or eclampsia, where increasing proteinuria is associated with maternal hypertension. In the past, serial quantitative assessments of proteinuria using dipsticks were used. Detection of significant proteinuria is now more usually carried out by measuring a protein to creatinine ratio in a random urine sample. A value of <30 g/mol helps to rule out significant proteinuria in hypertensive pregnant women. However, severity of the disease is not always related to absolute values.

Labour

The factors that lead to the onset of labour in humans appear to be complex, with a combination of biochemical and physical changes such as uterine stretch. In the first trimester, progesterone from the corpus luteum maintains uterine quiescence until the placenta takes over. In many species, it is falling concentrations of progesterone that initiate labour. This is not the case in humans, although inhibiting progesterone action does increase uterine activity and ripens the cervix. The uterus is primed for activity by a gradual rise in plasma oestrogen concentration and corticotrophin-releasing hormone (CRH) over a period of many weeks. The uterus then becomes sensitive to stimulants including oxytocics, prostaglandins and CRH. These appear to create a biochemical inflammatory-like reaction in the uterus.

Prostaglandins and oxytocin play central roles in human labour by stimulating myometrial contractility. Arachidonic acid is found in amniotic fluid and fetal membranes. This is converted by cyclooxygenase to prostaglandins. Clinical use of this process is seen with digital 'sweeping of the membranes' to enhance enzyme activity prior to therapeutic induction of labour, for which slow-release prostaglandin formulations are commonly used.

Oxytocin is a nonapeptide secreted by the posterior pituitary. Administration of oxytocin in pregnancy provokes uterine contractions and is used in the induction and acceleration of labour. Its exact role in the initiation and maintenance of normal human labour is not fully understood, particularly since normal parturition has occurred in women with oxytocin deficiency secondary to pituitary disease.

Corticotrophin-releasing hormone is a 41 amino acid peptide produced in the hypothalamus but in pregnancy it is also synthesized in the placenta and membranes. The concentration of free CRH rises markedly in the last three weeks of pregnancy. It promotes fetal cortisol production and prostaglandin production in the fetal membranes.

Pre-term labour (onset prior to 37 completed weeks of pregnancy) is the commonest cause of later disability and perinatal death. A major contributor to neonatal morbidity is lung dysfunction owing to lack of alveolar surfactant. Corticosteroid administration to the mother at least 24 h before delivery substantially improves outcome. Diagnosis of the onset of pre-term labour can be difficult but important, as tocolytics may delay delivery long enough for steroids to have an effect. An adjunct to assessing the likelihood of the onset of labour is the use of fetal fibronectin measurement. Fibronectins are large (450 kD) glycoproteins found in the cervical tissue, membranes and amniotic fluid. They act as an adhesive between the cells of the membranes and uterine lining and 'leak' out into the vagina if this is disrupted, as happens in pre-term labour. A bedside monoclonal antibody swab

test for fibronectin is now available, which is poor at predicting the onset of labour but has a negative predictive value of 99% that delivery will not occur within 14 days. Its use substantially reduces the need and cost for treatment or transfer of patients.

ORAL CONTRACEPTION AND HORMONE REPLACEMENT THERAPY

Introduction

Worldwide, many women are taking gonadal steroids (oestrogens and progestogens) as oral contraceptives (OCs) or sex hormone replacement therapy (HRT). Gonadal steroids have the potential to influence many metabolic systems and hence biochemical variables in the blood.

Methods for contraception can be divided into hormonal and non-hormonal methods. The latter include sterilization, barrier methods, the use of spermicides and natural methods. These do not have metabolic effects (although development of antisperm antibodies can affect the success of vasectomy reversal). Hormonal methods involve treatment of the woman (usually) with hormones to prevent conception, and have the potential to cause secondary metabolic disturbances.

Hormonal methods of contraception are summarized in Table 22.2. The most widely used are combined oral preparations of oestrogen and progestogen. Both components can have metabolic effects, in particular affecting lipid and glucose metabolism. Older combined oral contraceptives used relatively high doses of oestrogen, which could increase the concentration of thyroxine binding globulin, and hence increase total T4 (but not free T4) concentrations, but this is not a significant problem with modern combined oral contraceptives, in which lower doses of oestrogens are used.

There is an important difference between the oestrogen used in OCs and that used in HRT. If suppression

TABLE 22.2 Hormonal methods of contraception

Type	Components	Principle mode of action	Comments
Combined oral contraceptive pill	Oestrogen and progestogen	Inhibition of ovulation; also effects on cervical mucus and changes to endometrium	Typically taken for 21 days, followed by 7-day pill-free break
Combined contraceptive patch	Oestrogen and progestogen	As above	One patch worn for a week for three weeks, followed by one week with no patch
Combined contraceptive vaginal ring	Oestrogen and progestogen	As above	Worn for three weeks, followed by one week with no ring
Combined injectable contraceptive	Oestrogen and progestogen	As above	Monthly injections; not available in UK
Progestogen-only contraceptive pill	Progestogen	Multiple effects, including alteration of cervical mucus, suppression of ovulation and changes to endometrium	Taken daily with no breaks
Progestogen-only injectable contraceptive	Progestogen	As above	Injections at 2–3 monthly intervals according to product
Other progestogen-only devices	Progestogen	As above	Implants and intrauterine systems
Emergency contraception	Progestogen (levonorgestrel)	Probably prevents or delays ovulation	High dose of progestogen (50× that used in pill) but single dose and no metabolic sequelae

of ovulation is required, a synthetic oestrogen (almost always ethinyloestradiol) is necessary. In contrast, HRT requires lower doses of oestrogen, as the objective is to reverse the adverse effects of oestrogen deficiency. For this purpose, 'natural' (non-alkylated) oestrogens such as 17 β -oestradiol, oestradiol valerate and conjugated equine oestrogens have proved adequate, despite their ineffectiveness as contraceptive agents.

Contraceptive progestogens are used in order to increase contraceptive efficacy in combined regimens, on their own in the 'mini-pill' or to control menstrual bleeding. In contrast, HRT progestogens are used to prevent endometrial proliferation, hyperplasia and neoplasia. Again, the net dose of progestogen used in HRT is lower than that required for OC. Because natural progesterone is poorly absorbed when given orally, various synthetic progestogens have been developed. In the UK, the progestogens most commonly used in OC are levonorgestrel, norethisterone, gestodene and desogestrel, the latter two being novel derivatives chosen for their relatively low androgenicity. For postmenopausal HRT, levonorgestrel and norethisterone are most commonly used.

The majority of oral gonadal steroid therapies in current use combine an oestrogen with a progestogen. Progestogen-only oral contraceptives use very low doses of steroids and have comparatively little effect on plasma components.

Metabolic effects of oestrogens

Oestrogens suppress the secretion of FSH by the pituitary and hence suppress follicular maturation. The oestrogen in most combined oral contraceptives is ethinyloestradiol, at a daily dose between 15 and 50 μ g, but typically 30–35 μ g. Oestrogens tend to increase plasma triglyceride concentrations, reduce LDL-cholesterol and increase HDL-cholesterol concentrations. Oestrogens are prothrombotic, but their net effects on lipid metabolism tend to be anti-atherogenic. They have a negligible effect on glucose tolerance.

Metabolic effects of progestogens

The contraceptive effect of progestogens is related to suppression of LH secretion (tending to suppress ovulation) and to effects on cervical mucus that make it hostile to sperm, and on the endometrium that reduces the likelihood of successful implantation. Several different progestogens are used in combined oral contraceptives; they all lack the 19-methyl group that is present in progesterone and differ mainly (but not exclusively) in their C-17 substituents, whether or not they have a C-21 carbon and in the presence or absence of a methyl or ethyl group in the C-18 position. They include norethindrone, one of the first to become available for this purpose; norgestrel and levonorgestrel ('second-generation' progestogens), and the 'third generation' progestogens, desogestrel, gestodene and norgestimate. The metabolic effects of progestogens are largely dependent on their androgenicity. Levonorgestrel, gestodene, norgestimate and desogestrel (all 19-nor compounds, that is, lacking a 19-methyl group) have androgenic as well

as progestogenic activity, and tend to increase plasma LDL-cholesterol and reduce HDL-cholesterol concentrations. The effects of progestogens on glucose tolerance are considered below.

Metabolic effects of contraceptives

Effects of hormonal contraceptives on lipid metabolism and risk of vascular disease

The overall effect of combined oral contraceptives on lipid metabolism depends on the type and dose of oestrogen and progestogen and on the presence of any genetic tendency to hyperlipidaemia. In general, the effects of the oestrogen and progestogen tend to cancel each other out, though most modern agents tend slightly to reduce plasma LDL-cholesterol and increase HDL-cholesterol concentrations. They may also slightly increase triglyceride concentrations. Although it is not necessary to monitor plasma lipid concentrations in most women, this may be appropriate if there is known hyperlipidaemia, especially hypertriglyceridaemia.

It is noteworthy that in women who do not smoke cigarettes and who are normotensive, there is no evidence of an increase in the relative risk of cardiovascular disease associated with long-term use of combined oral contraceptives. Even in women who smoke or are hypertensive, the absolute risk, though higher, remains very low. The relative risk of ischaemic stroke is slightly increased, but the absolute risk remains very low. The risk of haemorrhagic stroke is unaffected in women below the age of 35, but slightly increased in older women.

The relative risk of venous thromboembolic disease is increased in users of combined oral contraceptives, although the absolute risk remains very low, and is certainly lower than that associated with pregnancy. The mechanisms involved are complex, but appear primarily to be the result of prothrombotic effects on the coagulation mechanism.

Contraindications to the use of the combined oral contraceptives include a personal history of stroke, myocardial infarction or venous thromboembolism, severe hypertension and heavy cigarette smoking.

There is no evidence of increased risk of venous thrombosis, cardiovascular or cerebrovascular disease in women using progestogen-only contraceptives.

Effects of oral contraceptives on glucose homoeostasis and diabetes

Any adverse effect of oral contraceptives on glucose tolerance appears to be related to the progestogen component. Progestogens antagonize the effects of insulin and increase the area under the curve of insulin response to oral glucose tolerance tests, but the effects on glucose tolerance, though statistically significant, are not clinically so.

There is no need to avoid combined preparations in patients with uncomplicated diabetes. Progestogen-only oral contraceptives are safe in patients with type 1 diabetes and can be recommended in many patients for whom combined products are contraindicated (e.g. heavy

smokers, severe hypertension). However, longer-acting progestogens, such as levonorgestrel, should be avoided in patients with type 2 diabetes.

Other metabolic effects of oral contraceptives

Hypertension may be exacerbated in OC users owing to an increase in the synthesis of angiotensinogen. However, most oral contraceptive users adapt their vasoactive controls and aldosterone production to compensate for the increase in blood pressure, and there is a poor correlation between the metabolic changes and actual changes in blood pressure.

Metabolic effects of injectable contraceptives

The metabolic effects of combined oral contraceptives in which the hormones are delivered other than orally are broadly similar to those seen with oral agents.

The metabolic effects of injectable progestogen contraceptives are minimal, with the exception that there is concern that depot medroxyprogesterone acetate may have an adverse effect on bone mineral density, probably related to a reduction in plasma oestradiol concentrations. The decrease in density appears reversible but whether there is still a later increased risk of osteoporosis is unknown and thus prolonged use, or use in women with other risk factors for osteoporosis, should be avoided. There is no evidence of significant adverse metabolic effects with progestogen-only implants or intrauterine delivery devices.

Although emergency contraceptives contain high doses of progestogens, exposure to the agent (levonorgestrel) is short-lived and has no adverse metabolic consequences.

Hormone replacement therapy

The primary aim of hormone replacement therapy in postmenopausal women is usually to reduce the clinical features of oestrogen deficiency, which include hot flushes, depression, vaginal dryness and consequent dyspareunia. Particularly in women who undergo a premature menopause, a secondary aim may be to reduce the risk of long-term consequences of oestrogen deficiency, particularly osteoporosis. In seeking to achieve these aims, it is clearly also important to avoid potentially harmful consequences.

Although menopausal symptoms are caused by oestrogen deficiency, oestrogen replacement should be accompanied by cyclical or continuous progestogen replacement in order to prevent endometrial hyperplasia: there is a clear association between the use of unopposed oestrogen treatment and the development of endometrial carcinoma. Women who have undergone hysterectomy can safely be treated with oestrogen alone.

Metabolic effects of the menopause

The onset of the menopause is accompanied by an increase in the plasma concentrations of LDL-cholesterol and triglycerides; HDL concentrations fall slightly, but with a rise in the HDL₃ fraction and a fall in the anti-atherogenic HDL₂ fraction. These changes are all

pro-atherogenic and are thought to be important factors in the increase in the prevalence of cardiovascular disease that occurs following the menopause.

Pancreatic insulin secretion tends to decline and, although there may be an increase in the half-life of insulin in the plasma, insulin sensitivity, and hence glucose tolerance, tend to decline. The distribution of body fat tends towards android rather than gynoid. Increases occur in the plasma concentrations of proteins involved in blood coagulation, including fibrinogen and antithrombin III, and there is a reduction in fibrinolytic activity. All these factors are likely to contribute to the increased risk of cardiovascular disease against a background of the increasing risk associated with advancing age.

Metabolic effects of HRT

The metabolic effects of HRT depend upon the steroids used, their doses and the route of administration, but in summary, oestrogen replacement tends to reduce LDL-cholesterol concentrations. Oestrogens also increase the clearance of remnant particles and reduce the concentration of lipoprotein(a). Although, when used alone, oestrogen replacement may increase HDL concentration, the effects of progestogens may counter this, depending on the androgenicity of the progestogen used. Non-androgenic agents may have little effect or even contribute to an increase in HDL-cholesterol. Conjugated equine oestrogens tend to increase plasma triglyceride concentrations, but oral oestradiol has a lesser effect and transdermal oestradiol reduces triglycerides.

Adverse consequences of hormone replacement therapy

Cyclical, combined oestrogen and progestogen replacement restores cyclical bleeding in the majority of women and is a common cause for them to abandon HRT. However, there are also possible long-term consequences. The most important of these include a small but significant increase in the risk of venous thromboembolism. There is no increase in the risk of endometrial cancer or ovarian cancer. The risk of breast cancer does increase in women who use combined HRT for more than five years, although only very slightly, but the effect declines after cessation of HRT and almost disappears after a further five years.

Hormone replacement therapy and heart disease

The increase in the risk of coronary heart disease following the menopause, together with the metabolic effects of HRT described above, suggested that HRT might be beneficial in the prevention of coronary disease. A major problem in adducing evidence to support this notion is the relatively small number of women who develop coronary disease in the years immediately following the menopause, necessitating the recruitment of large numbers of women into trials to provide the necessary statistical power. Although early epidemiological studies suggested that HRT reduces the risk of coronary disease,

meta-analyses of primary prevention trials suggest that HRT confers either no, or only a very small, degree of protection against cardiovascular death in younger women, and none in older women. The results of a large secondary prevention study showed no benefit. Indeed, one trial has indicated that HRT may cause an increase in cardiovascular events in older women in the first few years after starting treatment. The results of clinical trials thus appear to conflict with those of observational studies, and the reason for this remains uncertain. Thus HRT is not recommended for use solely for the primary or secondary prevention of coronary heart disease.

Hormone replacement therapy and osteoporosis

The increased rate of bone loss associated with oestrogen deficiency is a major cause of morbidity and mortality in older women (see Chapter 31); this is mainly due to the increased risk of hip fracture. Evidence from numerous trials is that HRT, whether with oestrogen alone or combined oestrogen and progestogen, increases bone mineral density and reduces the risk of hip fracture. However, although significant, the effect is small, and, except in women who sustain a premature menopause, HRT is not recommended solely for the prevention of fractures. Tibolone, a synthetic steroid with androgenic, oestrogenic and progestogenic properties, has also been shown to increase bone mineral density in postmenopausal women with osteoporosis.

ACKNOWLEDGEMENT

I would like to thank Irfana Koita-Kazi, John Waterstone, John H. Parsons, Mike Savvas and William J. Marshall who wrote this chapter for previous editions of the book.

Further reading

- Almog B, Shehata F, Suissa S et al. Age-related normograms of serum antimüllerian hormone levels in a population of infertile women: a multicenter study. *Fertil Steril* 2011;95(7):2359–63.
Paper on the use of anti-Müllerian hormone for evaluation of ovarian reserve and fertility.
- Balen AH, Franks S, Homburg R et al. Current management of polycystic ovary syndrome. London: RCOG Press; 2010.
Excellent synopsis of polycystic ovary syndrome.
- Balen AN, editor. Reproductive endocrinology for the MRCOG and beyond. London: RCOG Press; 2007.
This succinct text includes chapters on the endocrine changes that occur during puberty, the normal menstrual cycle, disorders of menstruation and hyperprolactinaemia.
- Bhattacharya S, Hamilton M, editors. Management of infertility for the MRCOG and beyond. London: RCOG Press; 2006.
A concise account of the management of infertility.
- Crook D, Godsland I. Safety evaluation of modern oral contraceptives. Effects on lipoprotein and carbohydrate metabolism. *Contraception* 1998;57:189–201.
A detailed analysis of the effects of a range of different formulations of oral contraceptives on lipid and glucose metabolism and their consequences.
- Johnson MH. Essential reproduction. 7th ed. Oxford: Blackwell; 2013.
A prizewinning text integrating biology, physiology and biochemistry of reproduction.

Lumsden MA, Rees M. Menopause and menopause transition. *Best Pract Res Clin Obstet Gynaecol* 2009;23:1–6.

Excellent volume. Includes cardiovascular and bone health in the menopause. National Institute for Health and Care Excellence (NICE). Antenatal care: routine care for the healthy pregnant woman. NICE Clinical Guideline 62; 2008. <http://publications.nice.org.uk/antenatal-care-cg62> [Accessed October 2013].

Homepage providing access to extensive information on antenatal care. (As of April 2013, NICE was re-named the National Institute for Health and Care Excellence.)

NHS Fetal anomaly screening program, <http://fetalanomaly.screening.nhs.uk/onlineresources>; [Accessed October 2013].

Online modules and training resources for screening for fetal anomalies and Down syndrome.

Stein IF, Leventhal ML. Amenorrhoea associated with bilateral polycystic ovaries. *Am J Obstet Gynecol* 1935;29:181–92.

The original description of the polycystic ovary syndrome: a classic text.

Vloeberghs V, Peeraer K, Pexsters A et al. Ovarian hyperstimulation syndrome and complications of ART. *Best Pract Res Clin Obstet Gynaecol* 2009;23(5):691–709.

Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy post-menopausal women: the Women's Health Initiative randomized controlled trial. *J Am Med Assoc* 2004;291:1701–12.

A study that suggested that HRT slightly increases the risk of coronary disease.

Langer RD, Manson JE, Allison MA. Have we come full circle – or moved forward? The Women's Health Initiative 10 years on. *Climacteric* 2012;15(3):206–12.

Answers many controversial aspects of the publication above.

APPENDIX 22.1 ACTH STIMULATION TEST FOR THE DIAGNOSIS OF CONGENITAL ADRENAL HYPERPLASIA

Perinatal diagnosis of congenital adrenal hyperplasia can be confirmed by a single assay of 17 α -hydroxyprogesterone (17-OHP) in a capillary blood sample. In less clear-cut cases, it may be necessary to measure concentrations of 17-OHP before and following stimulation. Consequently, 17-OHP concentrations are measured before and 60 min after an intravenous injection of 0.25 mg ACTH. The expected values 1 h after ACTH are as follows:

- unaffected adults 3–30 nmol/L
- heterozygotes for 21-hydroxylase deficiency 6–44 nmol/L
- patients with mild 21-hydroxylase deficiency 63–470 nmol/L

For further details of the diagnosis of heterozygotes and the late onset form of 21-hydroxylase deficiency, see New et al. (1983, below). A test assessing the ACTH and 17-OHP responses to corticotrophin releasing hormone (CRH) has been described that appears to distinguish classic from late onset congenital adrenal hyperplasia.

References

- Moreira AC, Elias LL. Pituitary-adrenal responses to corticotrophin-releasing hormone in different degrees of adrenal 21-hydroxylase deficiency. *Journal of Clinical Endocrinology and Metabolism* 1992;74:198–203.
- New MI, Lorenzen F, Lerner AJ et al. Genotyping steroid 21-hydroxylase deficiency: hormonal reference data. *Journal of Clinical Endocrinology and Metabolism* 1983;57:320–6.

Reproductive function in the male

John Miell • Zoe Davies

CHAPTER OUTLINE

INTRODUCTION 451

THE TESTES 451

Production and actions of testosterone 451

Hypothalamo–pituitary control of testicular function 452

Testicular malignancy 453

Endocrine disrupting chemicals 453

EVALUATION OF TESTICULAR FUNCTION 454

Semen analysis 454

Endocrine evaluation: hypothalamo–pituitary–gonadal axis 454

MALE HYPOGONADISM 455

Clinical features 455

Primary hypogonadism 455

Secondary hypogonadism 455

Defective hormone synthesis and hormone receptor defects 456

Treatment of hypogonadism 456

GYNAECOMASTIA 457

Causes of gynaecomastia 457

Investigation 458

IMPOTENCE 458

Investigation 458

Treatment of erectile impotence 459

APPENDIX 460

INTRODUCTION

Normal male reproductive function – the manufacturing of semen, the ability to create and sustain penile erection and to ejaculate – is dependent on the physiological integration of the hypothalamo–pituitary–testicular axis. Production of androgens, normal responses in the tissues on which these hormones act and intact neurological pathways required for ejaculation are crucial for full operation of this system, to make human reproduction possible.

THE TESTES

The testes have two distinct roles: androgen production and spermatogenesis. These are achieved by two main units – the interstitial cells and the seminiferous tubules – under the control of the hypothalamo–pituitary axis.

Production and actions of testosterone

Testosterone is the major androgen produced in the interstitial cells of the testes. Also known as Leydig cells, these are found in clumps between the seminiferous tubules, and their mass correlates positively with androgen production. Sertoli cells, along with germ cells, form the seminiferous tubules (Fig. 23.1) and secrete

hormones that control embryological sexual differentiation. In addition, they support spermatogenesis by regulating the composition of the seminiferous tubular fluid, thereby providing an environment for meiotic germ cell development.

At week eight of gestation, a human fetus has both Müllerian and Wolffian ducts, with the potential to develop into female and male genital tracts, respectively, and differing only by its sex chromosomes. The Y chromosome is a powerful sex determinant – it confers maleness to the extent that even XXXXY fetuses are phenotypically male at birth. (This topic is covered in more detail in Chapter 21.)

Normal male sexual differentiation in utero depends on the presence of testosterone and other hormones produced by these embryological gonads. Leydig cells are stimulated by placental human chorionic gonadotrophin (hCG) to secrete testosterone from the ninth week of gestation, leading to development of the Wolffian ducts. Sertoli cells secrete anti-Müllerian hormone (AMH), which causes regression of the Müllerian ducts and oögonia (the precursors of primary oocytes). In the absence of AMH and testosterone, the Müllerian ducts differentiate into female internal genitalia. Lower concentrations of testosterone are maintained during late gestation by secretion of luteinizing hormone (LH) from the fetal pituitary. The presence of 5 α -reductase, and the subsequent production of dihydrotestosterone (DHT) from testosterone,

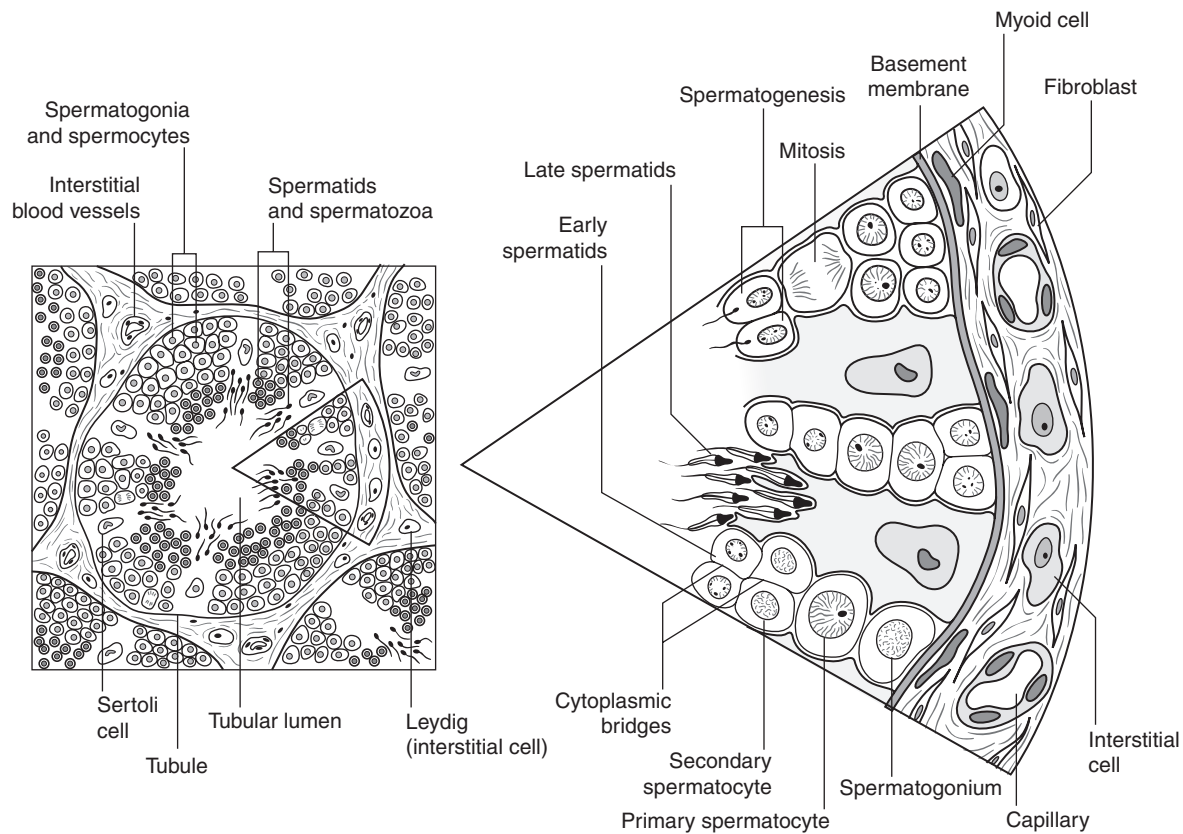


FIGURE 23.1 ■ The seminiferous tubules.

is required for development of the urogenital sinus. Testosterone secretion may continue for many months post partum before declining until the onset of puberty, when it will again be essential for sexual development.

Testosterone secretion from Leydig cells follows stimulation by LH, which binds to specific membrane receptors. Luteinizing hormone activates adenylyl cyclase, resulting in formation of cAMP, which binds to protein kinase and in turn activates the enzyme 20,22-desmolase, which is responsible for conversion of cholesterol to pregnenolone (see Fig. 22.5, p. 438). The synthesis of testosterone is completed by four other enzymes (3β -hydroxysteroid dehydrogenase, 17α -hydroxylase, $17,20$ -desmolase and 17β -hydroxysteroid dehydrogenase). At the target organ, testosterone is transported to the nucleus, and gene activation results in the formation of specific messenger RNAs and proteins. Testosterone is usually converted to the active metabolite DHT by the enzyme 5α -reductase. In the adult, there is circadian variation in testosterone secretion with a peak in the early morning and a trough at about 18.00 h, and therefore blood samples for testosterone measurement should ideally be taken at 09.00 h. The biological actions of testosterone and DHT are summarized in Table 23.1.

Hypothalamo–pituitary control of testicular function

Gonadotrophin releasing hormone (GnRH) is secreted from the hypothalamus in a pulsatile fashion, with peaks

TABLE 23.1 Biological actions of testosterone and dihydrotestosterone

	Testosterone	Dihydrotestosterone
Fetus	Wolffian stimulation	Differentiation of external genitalia Growth of phallus
Puberty	Musculoskeletal development	Full secondary sexual hair Temporal recession of head hair Prostate growth
Adult	Cessation of growth in long bones Sebum secretion Spermatogenesis Libido	Seminal vesicle growth Prostatic hypertrophy Male pattern baldness

occurring approximately every 90 min. It regulates the release of LH, which is crucial for the regulation of Leydig cell number and function, and hence production of testosterone, and follicle stimulating hormone (FSH), which stimulates Sertoli cell division and growth and is therefore responsible for testicular enlargement at puberty. Faster pulse frequencies favour LH release and slower ones, FSH. The pulsatility of GnRH secretion is important in the regulation of gonadotrophin secretion, as continuous exposure results in down-regulation of the GnRH receptor. Luteinizing hormone secretion is also highly pulsatile, which is reflected in measured

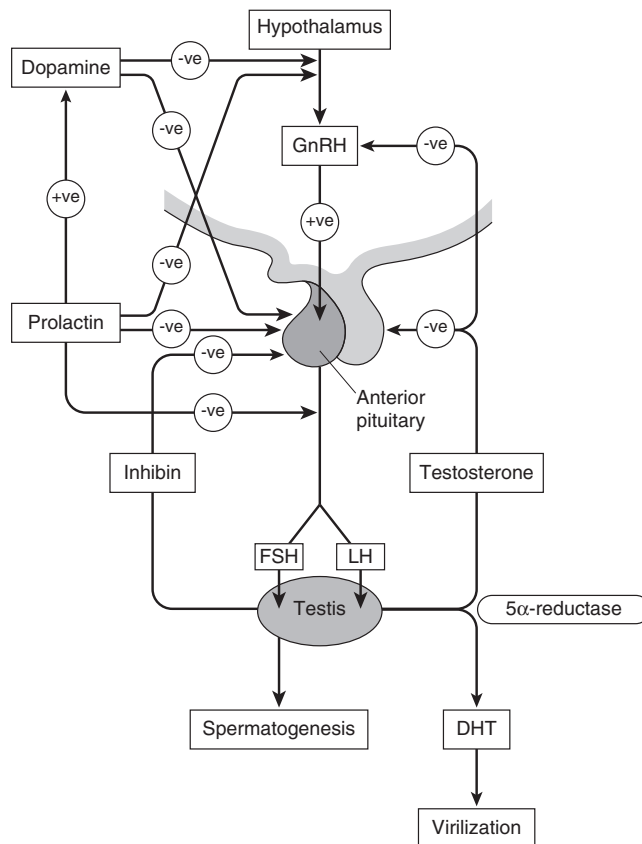


FIGURE 23.2 ■ Hypothalamo-pituitary control of testicular function.

plasma concentrations. Plasma concentrations of FSH are less variable, presumably as a result of slower metabolic clearance.

The gonadotrophins are, in turn, regulated by a series of feedback mechanisms (Fig. 23.2). Testosterone inhibits LH and, to a lesser extent, FSH secretion from the anterior pituitary. Oestradiol, which is formed in the male through metabolism of androgens by aromatase found in adipose tissue, skin, kidneys and brain, in addition to Leydig cells, also inhibits gonadotrophin secretion by the pituitary. Opioids and prolactin reduce the pulsatile activity of GnRH and thus decrease FSH and LH secretion; prolactin also inhibits the direct action of gonadotrophins on the testes.

Follicle stimulating hormone is also regulated by inhibin B, which is produced by Sertoli cells and inhibits FSH production to regulate spermatogenesis. Conversely, activin works to enhance FSH action in the gonads, increasing spermatogenesis. Both are secreted in the seminal fluid and are thought to work locally in addition to centrally on the pituitary. Inhibin B can be used as a marker for male factor infertility as concentrations directly reflect the function of the Sertoli cells: plasma inhibin B concentrations are significantly higher in fertile men than in infertile men.

Gonadotrophin secretion is affected by a diverse range of regulators, including stress and nutritional status. Recent studies have shown that low concentrations of leptin, an adipose-derived hormone, or decreased

sensitivity to leptin, such as occurs in obese individuals, are associated with decreased gonadotrophin concentrations and decreased fertility. This is mediated via hypothalamic kisspeptin neurons rather than directly via GnRH neurons. Kisspeptin is a neurohormone thought to have a key role in puberty, although this is yet to be clearly defined. Patients with mutations in the kisspeptin gene fail to go through puberty owing to hypogonadotropic hypogonadism. In animal models, treatment with kisspeptin can restore GnRH pulsatility and gonadotrophin secretion.

Testicular malignancy

Testicular cancer is the commonest malignancy of men aged 20–39. Histologically, it is commonly a germ cell malignancy – either a seminoma or, more often, a teratoma. Patients are likely to present with a testicular mass, and as this cancer is relatively easily detectable with regular self-inspection/examination, and has a high cure rate if detected prior to metastatic spread, there have been several public campaigns to raise awareness amongst young men.

Approximately 80% of testicular teratomas produce hCG or α -fetoprotein (AFP) (see Chapter 42). Human chorionic gonadotrophin is very similar to LH in structure. It stimulates the LH receptor in the testis eventually resulting in increased secretion of oestradiol, and hence gynaecomastia. Some rarer testicular malignancies, such as Sertoli cell tumours, can also secrete oestradiol directly. Human chorionic gonadotrophin can interact with the thyroid hormone axis as an agonist of the thyroid stimulating hormone (TSH) receptor, to cause hyperthyroidism.

Endocrine disrupting chemicals

Endocrine disruptors are environmental substances that interfere with normal hormonal systems. They are very diverse and can be plastics, chemicals, pesticides and fungicides. They are often able to interact with steroid hormone receptors and can act as androgens and anti-androgens. Exposure may be occupational, but also through water supply, soil and food. There may be a lag between exposure and clinical presentation and there may be trans-generational effects if mutations or modifications of gene expression occur. The developmental stage at which a person is exposed may be crucial. The observations of a declining sperm count, and increasing rates of testicular cancer throughout the last century prompted researchers to ask if these epidemiological trends could be explained by endocrine disruptors. In animal models, exposure to phthalates is significantly related to oligospermia, and epidemiological evidence suggests this may be the case for humans. Endocrine disruptors have also been linked to cryptorchidism, hypospadias, early or delayed puberty and prostatic hyperplasia. More detail on this topic is beyond the scope of this chapter, but the interested reader will find more material in the further reading section, below.

EVALUATION OF TESTICULAR FUNCTION

The laboratory evaluation of testicular function may be conveniently divided into assessment of spermatogenesis and assessment of endocrine gonadal function.

Semen analysis

Semen analysis is used both in the evaluation of male fertility and the follow-up of treatment regimens for male subfertility. It must be noted that there is wide inter-laboratory variation in the results, although this has been reduced in recent years by automated systems using electro-optics or computer-assisted analysis. Furthermore, there is marked variation in sperm output on a day-to-day basis. It is often more useful in clinical practice to assess the actual fertilization capacity of sperm *in vitro*.

A semen sample is collected by masturbation and should ideally be analysed within 1 h. Basic semen analysis measures the number of spermatozoa (per unit volume and per ejaculate), motility and morphology. These are assessed according to World Health Organization (WHO) criteria, updated in 2010, although it should be noted that individual laboratories will also have developed their own reference ranges. The WHO reference ranges were determined by the analysis of semen from fertile men whose partners became pregnant within one year of trying to conceive. Average sperm count is >60 million/mL and a count of <15 million/mL is considered subfertile. There should be >39 million sperm in a given ejaculate and semen volume should be 1.5 mL or greater.

Motility is defined as the percentage of moving sperm and is subdivided into the following categories:

- progressive motility (PR): spermatozoa moving actively, either linearly or in a large circle, regardless of speed
- non-progressive motility (NP): all other patterns of motility with an absence of progression
- immotility (IM): no movement.

The WHO reference values state that $\geq 40\%$ of the sperm should exhibit either PR or NP motility and $\geq 32\%$ should exhibit PR motility. At least 58% of sperm should be alive – this is especially important for samples with a PR motility of <40% and must be assessed within 1 h of ejaculation.

Morphology, describing defects in the appearance of spermatozoa, is most commonly assessed according to the WHO morphology standards. The WHO lower reference limit for normal morphology is 4%. Few samples exhibit >25% normal spermatozoa.

Other parameters measured include volume, pH, white blood cell count, antisperm antibodies, colour, viscosity, agglutination, aggregation and time until liquefaction. Reference ranges for semen analysis are shown in [Table 23.2](#).

Endocrine evaluation: hypothalamo–pituitary–gonadal axis

Determination of basal plasma LH and FSH concentrations is a routine part of the assessment of the hypothalamo–pituitary–gonadal axis. As LH and, to a

TABLE 23.2 Reference ranges for semen variables^a

Variable	Reference range
Volume	>1.5 mL
pH	>7.2
Sperm concentration	>15 × 10 ⁶ sperm/mL
Total count	>39 × 10 ⁶ sperm/ejaculate
Motility	40% or more with total motility (PR + NP)
	32% or more with PR motility
Vitality	58% or more alive
White blood cells	<1 × 10 ⁶ cells/ejaculate
Fructose	>13 μmol/ejaculate

^aWHO classification described in text. PR, progressive motility; NP, non-progressive motility.

lesser extent, FSH are secreted in a pulsatile fashion, more accurate interpretation may be possible if serial samples are taken over a period of time. However, in routine clinical practice, the measurement of basal plasma concentrations of LH and FSH with testosterone provides most of the information required.

Elevated LH and FSH in the presence of low testosterone concentrations suggest primary testicular failure; low or normal LH and FSH associated with low testosterone concentrations suggest a hypothalamo–pituitary disorder. It is important to recognize that the latter may be caused by any acute illness. Rarely, LH alone may be elevated with a low testosterone (the Sertoli-only syndrome) and, in patients with abnormal spermatogenesis, FSH may be raised with a normal testosterone. It should be remembered that pituitary and gonadal dysfunction may coexist, as in haemochromatosis. In this situation, LH, FSH and testosterone concentrations are low.

The GnRH test ([Appendix 23.1i](#)) has been used to assess pituitary reserve of LH and FSH and overall hypothalamo–pituitary function. This test is now used relatively infrequently, as it rarely provides extra information beyond that from basal LH and FSH concentrations.

The clomifene stimulation test ([Appendix 23.1ii](#)) can be used to assess the hypothalamo–pituitary axis with respect to gonadotrophin secretion. Clomifene competitively inhibits hypothalamo–pituitary androgen binding. The normal response is a doubling of LH and FSH by day 10 after administration. A low response, however, does not distinguish between abnormalities at the pituitary or hypothalamic level.

Plasma testosterone concentrations exhibit diurnal variation throughout the day – some studies suggesting a variation of as much as 30% between a sample taken at 08.00 h and one at 16.00 h. Consequently, and by convention, testosterone is usually determined on a sample taken before 09.00 h. Analysis of serum testosterone concentrations must take into account sex hormone-binding globulin (SHBG), as 50% of circulating testosterone is bound to SHBG. Low testosterone concentrations are seen both in primary testicular disease, indicating poor Leydig cell function, and hypothalamo–pituitary disorders with

depressed gonadotrophin secretion. Elevated concentrations are seen in patients with gonadotrophin- or androgen-secreting tumours (e.g. Leydig cell tumours), in whom gonadotrophin concentrations will be low.

The hCG stimulation test (Appendix 23.1iii) can be helpful in differentiating primary testicular failure from gonadotrophin deficiency. Failure of testosterone to rise after administration of hCG suggests inadequate Leydig cell function. An exaggerated response suggests secondary hypogonadism. The test is also useful in confirming the presence of testicular tissue in cryptorchidism and in conditions in which there may be combined pituitary and testicular dysfunction, for example as a result of iron deposition in haemochromatosis or treated thalassaemia.

MALE HYPOGONADISM

Male hypogonadism may result from either primary testicular failure, in which the testes fail to develop properly or are injured by disease or medical manipulation, or secondary testicular failure owing to hypothalamic or pituitary lesions (hypogonadotrophic hypogonadism). The latter may result from developmental or acquired pituitary disease or a failure of hypothalamic GnRH secretion. Other syndromes resulting in hypogonadism include enzyme defects in which hormonal synthesis is interrupted despite an intact hypothalamo-pituitary-testicular axis.

Clinical features

The physiological effects of testosterone are listed in Table 23.1. Prepubertal testicular failure results in lack of sexual maturation with persistent infantile genitalia and absence of pubic and axillary hair growth. Delayed epiphyseal closure leads to the development of a eunuchoid habitus with increased arm span (>5 cm greater than height) and increased lower body height (sole-to-pubic symphysis length >5 cm greater than pubic symphysis-to-head length). The reduction in testosterone is also associated with diminished anabolism and a decrease in normal male muscular development.

Postpubertal onset of androgen deficiency is less obvious and usually presents with decreased libido, impotence and infertility. There may be an associated decrease in growth of facial, pubic and axillary hair and a diminution in skeletal musculature. Gynaecomastia may occur as a result of an increase in the oestradiol:testosterone ratio (see below). Despite some reduction in testicular volume, the penis and prostate do not shrink and erections and orgasms may continue. Nevertheless, spermatogenesis requires testosterone and, even when able to ejaculate, severely hypogonadal males are azoospermic.

Primary hypogonadism

This implies primary testicular failure and is characterized by low plasma testosterone and elevated LH and FSH concentrations. The chief causes can be divided into genetic abnormalities and cryptorchidism. Other causes are

orchitis, chronic illness, drugs (including excess alcohol), chemo- and radiotherapy and, rarely, testicular trauma.

Genetic causes

In the male, the most common of these is Klinefelter syndrome, thought to occur in 1 in 500 live births. This was described in 1942 by Klinefelter, Reifenstein and Albright, and comprises eunuchoid stature, gynaecomastia, small testes and infertility. It is caused by non-disjunction of the sex chromosomes during meiosis, the most common resulting abnormality being 46XXY but occasionally other X states or mosaicism (e.g. 48XXXY/46XY) are seen. All these cause Leydig cell abnormalities, resulting in impaired testosterone secretion and seminiferous tubular function, and decreased inhibin secretion. Other genetic causes of primary hypogonadism include XX males, Noonan syndrome, XY/XO gonadal dysgenesis and deletion of the Y chromosome in part or entirety.

Cryptorchidism

In 1–2% of male neonates, the testes fail to descend bilaterally from the abdominal cavity and do not produce adequate concentrations of testosterone. In addition, there is an increased risk of testicular malignancy. Surgery to position the testes in the scrotum, particularly if carried out in early childhood, reduces the risk of later malignancy, and facilitates surveillance. For this reason, it is important to distinguish this condition from anorchia, complete absence of testes in a 46XY male. The presence or absence of gonadal tissue can be assessed using ultrasound or magnetic resonance imaging scanning. If functional testicular tissue is present, there will be an increase in testosterone during an hCG stimulation test.

Secondary hypogonadism

This is hypogonadism caused by hypothalamic or pituitary dysfunction, resulting in deficient gonadotrophin secretion and leading to a low testosterone concentration. The causes may be divided into congenital and acquired disorders.

Congenital causes

The most common of these is Kallmann syndrome. Mutations or deletions of the *KAL1* gene, which encodes anosmin-1, result in the X-linked form, leading to an abnormal migration of GnRH-producing neurons to the hypothalamus and failure of episodic GnRH secretion. The syndrome manifests as scant secondary sexual development and eunuchoid habitus in association with anosmia and, occasionally, other craniofacial abnormalities such as cleft lip and palate.

Congenital adrenal hypoplasia presents with primary adrenal failure in association with inadequate GnRH secretion. It is caused by a mutation of the *DAX-1* gene on the X chromosome. The fertile eunuch syndrome is a selective LH deficiency leading to partial virilization with normal Sertoli cell function. This can be treated with testosterone replacement or hCG.

Homozygous mutations of the leptin gene and receptor have been found to be associated with idiopathic hypogonadotrophic hypogonadism (IHH) and severe obesity as a result of impaired GnRH release. Idiopathic hypogonadotrophic hypogonadism has also been associated with mutations of the FSH β -subunit, LH β -subunit, *PC1* (prohormone convertase 1) and *GPR54* (part of the kisspeptin pathway) genes. Other rare congenital causes of hypogonadotrophic hypogonadism include Prader–Willi syndrome and Laurence–Moon–Biedl syndrome. Haemochromatosis is another inherited condition sometimes associated with the development of secondary hypogonadotrophic hypogonadism, although the presentation of this complication is usually rather later in life.

Acquired causes

There are a number of acquired conditions leading to reduced gonadotrophin secretion (Box 23.1). These include excessive exercise or weight loss, physical or psychological stress and systemic illness. Direct damage to the hypothalamus or pituitary, such as from head trauma, radiotherapy, surgery, pituitary tumours and infiltrative disorders (e.g. sarcoidosis) can also be a cause, in addition to prescribed medications causing hyperprolactinaemia and illicit drugs such as cocaine, anabolic steroids and opioids.

Defective hormone synthesis and hormone receptor defects

These conditions lead to ambiguous external genitalia with a 46XY genotype. There are rudimentary or absent Wolffian duct derivatives (seminal vesicles and ductus deferens), but because of the presence of AMH, the uterus and fallopian tubes do not develop either. In addition to this there is usually cryptorchidism.

Five enzymes are responsible for the conversion of cholesterol to testosterone; defects in each of these enzymes have been reported (see Chapter 21). Plasma testosterone concentrations are low and hCG stimulation results in elevated concentrations of precursors proximal to the enzyme block. Defects in androgen receptor

proteins can also lead to hypogonadism, but with normal testosterone concentrations.

5 α -Reductase deficiency

5 α -Reductase converts testosterone to dihydrotestosterone (DHT). Dihydrotestosterone is responsible for masculinization, and normal plasma concentrations are a prerequisite for the full development of the phallus, scrotum and prostate. In 5 α -reductase deficiency, external genitalia at birth are ambiguous with a urogenital sinus, clitoromegaly, bifid scrotum, blind vaginal pouch and inguinal or labial testes. Wolffian duct derivatives are present, as they develop under the influence of testosterone rather than DHT. The deficiency is caused by a mutation in the 5 α -reductase type 2 gene, which is inherited in an autosomal recessive fashion.

At puberty, patients undergo virilization with some phallic enlargement, development of male habitus and psychosexual orientation and descent of the testes, occasionally with spermatogenesis. This occurs either as a result of a massive increase in testosterone secretion at puberty or by conversion of some testosterone to DHT if the enzyme deficiency is incomplete. Patients may respond to very high doses of testosterone or to DHT (the latter can be applied locally as a 2% cream with some beneficial effects on penile length). If peripubertal virilization is minimal, patients may be raised as infertile females after removal of the cryptorchid testes.

Androgen insensitivity syndromes

These are caused by absence or deficiency of the DHT receptor protein. The complete form results in an apparently normal female phenotype with absent pubic and axillary hair and is transmitted as an X-linked recessive trait. Administration of testosterone has no effect on the generation of secondary sexual hair. As patients usually have normal female psychosexual orientation, treatment involves removal of cryptorchid testes and subsequent oestrogen replacement. Normal female sexual activity is possible, although, of course, the patients remain infertile.

In partial androgen insensitivity syndrome, presentation varies according to the degree of DHT receptor deficiency. It ranges from apparent clitoromegaly in a phenotypically female neonate at birth, to infertility in an apparently normal adult male. Diagnosis is confirmed by identification of the androgen receptor gene mutation. The management of these patients depends on age at diagnosis and the response to treatment with exogenous testosterone.

Treatment of hypogonadism

Testosterone replacement remains the mainstay of treatment. Traditionally, this has been administered by implantation of subcutaneous pellets or by long-acting mixtures of testosterone esters (e.g. Sustanon[®]). The latter are given by deep intramuscular injection every 2–4 weeks. More recently, patches have been used, although there have been problems with varying degrees of absorption resulting in unpredictable plasma testosterone concentrations,

BOX 23.1 Acquired causes of hypogonadotrophic hypogonadism

- Malnutrition and weight loss
- Excessive exercise
- Stress
- Chronic illness
- Suprasellar tumours
 - Craniopharyngiomas
 - Pinealomas
- Destructive pituitary disease
 - Pituitary adenomas
 - Trauma, surgery or radiotherapy
 - Empty sella syndrome
- Sarcoidosis
- Prolactinomas
- Iron overload
 - Frequent blood transfusion, haemochromatosis
- Oestrogen-secreting adrenal tumours
- Drugs

and also problems with adhesion and hypersensitivity at the site of application. Gel and buccal mucoadhesive preparations are now available and are widely used. These are more convenient and result in much more even plasma concentrations, although a disadvantage of gel preparations is inadvertent transfer of gel to women or children. Oral preparations have been used but are poorly absorbed. As the prostate and seminal vesicles are androgen sensitive, there is a risk of stimulating growth of a prostatic carcinoma and therefore plasma prostate specific antigen concentrations should be monitored. (Treatment regimens are discussed in more detail in Chapter 21.)

Testosterone replacement will not achieve fertility in primary testicular failure. Only in secondary hypogonadism can infertility be treated. Treatment in hypogonadotropic hypogonadism is with intramuscular hCG with the addition of human menopausal gonadotrophin to supply FSH activity as necessary. This will usually induce a good quality ejaculate, which is potentially capable of resulting in conception despite a relatively low sperm count (1–20 million/mL). An alternative treatment is with small pulses of GnRH delivered by a programmed pump, mimicking physiological GnRH pulsatility. These are not long-term treatment options, partly because hCG, whether endogenous or exogenous, causes gynaecomastia. When the need for fertility is over, patients should resume treatment with testosterone.

Fertility can also be achieved using assisted reproduction techniques; in male infertility these include intrauterine insemination and intracytoplasmic sperm injection (ICSI). In ICSI, spermatozoa are injected directly into an oocyte and the fertilized embryo is then inserted into the uterus. ICSI can be appropriate where there is a very low sperm count, poor morphology or motility of sperm. It is also useful where there is an interruption in the pathway from testes to urethra such as occurs post-vasectomy or with an absent vas deferens in patients with cystic fibrosis. Abnormal sperm morphology does not appear to have a negative effect on blastocyst formation and birth defect rates are not significantly different from those conceived using conventional in vitro fertilization. However, as there is some evidence of an increased number of chromosomal abnormalities in children conceived using ICSI, concerns remain regarding health and fertility later in life.

Plasma testosterone concentrations gradually decline with age after the third decade, as a result of a number of factors including decreased GnRH release, increased sensitivity to the negative feedback of testosterone and reduced responsiveness of Leydig cells. Symptoms associated with testosterone deficiency, such as reduced exercise tolerance and increased body fat, overlap with changes linked to the natural process of ageing and hence there is an unavoidable implication that testosterone replacement in the elderly will be beneficial. Positive effects of replacement include increased bone density, increased lean body mass, improvement of lipid profile and stimulation of red blood cell production. There is a concern that such replacement may be associated with an increased risk of prostatic carcinoma, but this has not been proven.

Although testosterone replacement may be indicated in elderly patients with clear testosterone deficiency and no contraindications to treatment, the approach to marginally low concentrations is controversial. Initiation of treatment

should always be preceded by evaluation to rule out pre-existing contraindications, for example prostatic carcinoma or polycythaemia, and followed by regular monitoring.

GYNAECOMASTIA

Excessive development of the male mammary glands with increases in both stromal and glandular tissue (gynaecomastia) accounts for ~70% of male breast disorders. In the male, breast development may vary from a small subareolar button of tissue to florid breast development with feminization of the nipples and associated breast tenderness.

Gynaecomastia may occur in neonatal life, when it is usually transient and caused by transplacental transfer of maternal oestrogens into the fetal bloodstream. During puberty, gynaecomastia of some degree occurs in up to 60% of adolescent boys: this may reflect normal physiological development and is typically followed by spontaneous regression after 12–18 months.

Causes of gynaecomastia

Physiological gynaecomastia during puberty probably occurs as a result of the relative changes in oestradiol and testosterone concentrations – in normal early puberty, oestradiol concentrations are high in comparison with those of testosterone and this increased oestradiol:testosterone ratio can induce breast development.

Indeed, pathological causes are invariably related to imbalances between testosterone and oestrogen synthesis and action, and can be broadly divided into hypogonadism, other endocrine disorders, tumour related, systemic disease (Box 23.2) and drugs (Table 23.3).

BOX 23.2 Causes of gynaecomastia

Hypogonadism

- Primary
 - Castration
 - Orchitis
 - Klinefelter syndrome
 - Cryptorchidism
- Secondary
 - Prolactinoma (effect of tumour expansion or prolactin)

Other endocrine disorders

- Thyrotoxicosis
- Cushing syndrome
- Congenital adrenal hyperplasia
- Androgen insensitivity

Tumours

- Testicular (Leydig cell, Sertoli cell, choriocarcinoma)
- Adrenal (adenoma and carcinoma)
- Bronchogenic carcinoma
- Hepatocellular carcinoma
- Others – gastric, renal, haematological

Systemic disease

- Chronic kidney disease
- Liver failure
- Haemochromatosis

TABLE 23.3 Drugs implicated in the pathogenesis of gynaecomastia

Drug	Mechanism
Oestrogens	Direct stimulation, inhibition of androgen production
Androgens	Aromatized to oestrogens
Digitalis	Binding to oestrogen receptors
Tetrahydrocannabinol	Binding to oestrogen receptors
Griseofulvin	Binding to oestrogen receptors
Ketoconazole	Increase in oestradiol:testosterone ratio
Spirololactone	Anti-androgen
Cimetidine	Anti-androgen
Cyproterone	Anti-androgen
Methadone	Mechanism unknown
Phenothiazines	
Reserpine	
Isoniazid	

Investigation

A full drug history and a careful physical examination are essential, with particular reference to the testes, thyroid, liver and lungs.

A suggested schedule for the investigation of symptomatic gynaecomastia is illustrated in [Figure 23.3](#) and the interpretation of endocrine tests in [Table 23.4](#).

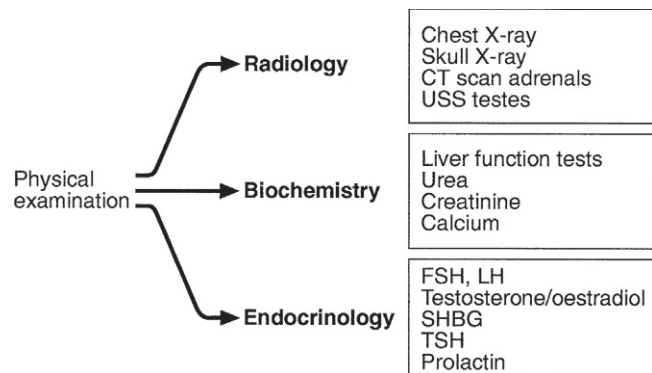
IMPOTENCE

Erectile impotence is defined as the inability to achieve and maintain a penile erection of sufficient quality to permit penetrative intercourse. In practice, the commonest physical cause of impotence is diabetes mellitus – some 50% of diabetic males will become impotent after the age of 50. Both autonomic neuropathy and vascular insufficiency may contribute to this. Other endocrine causes include hypogonadism (primary or secondary), hyperprolactinaemia and thyroid disease. Other causes are summarized in [Box 23.3](#).

A psychogenic aetiology is suggested by a history of early morning or sleep erections or full erection on masturbation or with other partners. Treatment with certain drugs can result in impotence – these are summarized in [Box 23.4](#).

Investigation

The initial history and examination is of great importance. Reduced frequency of shaving, reduced libido,

**FIGURE 23.3** Investigation of gynaecomastia. CT, computerized tomography; USS, ultrasound scan.

BOX 23.3 Causes of erectile impotence

- Diabetes mellitus
- Psychogenic
 - Depression
 - Anxiety
- Endocrine
 - Hyperprolactinaemia
 - Thyroid disease (hyper- and hypothyroidism)
 - Hypogonadism
- Peyronie disease
- Alcohol abuse
- Spinal cord damage (neurogenic)
 - Multiple sclerosis
 - Surgery
 - Trauma
- Vascular disease
- Chronic systemic disease
 - Chronic kidney disease
 - Disseminated malignancy
 - Liver disease
- Drugs – see [Box 23.4](#)

small testes and regression or lack of secondary sexual characteristics suggests hypogonadism. Testicular tumours may be palpable and visual field assessment and imaging of the brain and skull should be carried out if there is a possibility of pituitary tumour. A full neurological and cardiovascular examination is essential. The relevant basic biochemical investigations are listed in [Box 23.5](#).

TABLE 23.4 Interpretation of endocrine tests in the investigation of gynaecomastia (only those changes important in diagnosis are shown)

Condition	Hormone						
	Testosterone	LH	FSH	SHBG	Oestradiol	TSH	hCG
Primary testicular failure	↓	↑	↑				
Secondary testicular failure	↓	↓	↓				
Primary oestrogen-secreting tumour	↓	↓			↑		
hCG-secreting tumour		↑ ^a			↑		↑
Chronic liver disease	↓	↓	↓	↑			
Thyrotoxicosis	↑	↑	↑	↑		↓	

^aDepends on the specificity of the LH assay; modern assays show very little cross-reactivity with hCG.

BOX 23.4 Substances causing erectile impotence

- Oestrogens
- Cyproterone
- Dopamine antagonists
- Cimetidine
- Spironolactone
- Alcohol
- Cigarette smoke
- Guanethidine
- Methyldopa
- Clonidine
- Thiazide diuretics
- β -Blockers
- Disopyramide

BOX 23.5 Investigation of erectile impotence**Biochemical tests**

- Serum glucose and lipids
- Renal function
- Testosterone
- Thyroid function
- LH, FSH
- Prolactin
- Others as indicated clinically

Tests for vascular causes

- Duplex ultrasonography
- Papaverine test
- Pudendal arteriography
- Nocturnal tumescence testing
- Penile brachial index

Tests for neurogenic causes

- Pudendal nerve-evoked potentials
- Biothesiometry

Previously, treatment options were invasive and hence thorough justification by extensive investigation was necessary. Following a dramatic improvement in the treatment options in the last two decades, owing mainly to the availability of phosphodiesterase (PDE)-5 inhibitors, more extensive specialist evaluation of vascular and neurological causes (Box 23.5) is now reserved for a limited subset of patients.

For example, penile arterial blood pressure may be investigated. Penile systolic blood pressure is measured with a Doppler stethoscope and recorded as a ratio to that of the brachial artery. Low penile:brachial ratios (<0.6) suggest an arterial cause for impotence. Continuous arterial blood pressure recordings can be achieved using a small cuff around the base of the penis attached via a transducer to a recorder. If the arterial supply is impaired, the recording will show small amplitude pulses with abnormal waveforms.

Normal men experience nocturnal erections during rapid eye movement (REM) sleep and their absence suggests an organic cause for impotence. Nocturnal tumescence testing is ideally carried out in a sleep laboratory and should be repeated on three or more occasions to maximize

diagnostic accuracy. Other techniques include pudendal arteriography and duplex ultrasonography, but a detailed discussion of these is beyond the scope of this book.

Techniques for the investigation of neurological causes include biothesiometry, which is a simple test of afferent sensory function. An electromagnetic test probe is placed on the penile skin in various sites and the intensity of vibration increased until it is perceived by the patient. The degree of vibration required for perception can be compared with other sites on the patient's body.

Treatment of erectile impotence

Treatment options for erectile impotence include oral therapy in the form of PDE-5 inhibitors (sildenafil, vardenafil, and tadalafil), intraurethral alprostadil, intracavernosal injection of vasodilators, vacuum devices and surgery (bypass and implants).

The advent of the PDE-5 inhibitors has greatly simplified the treatment of impotence. The available agents have similar efficacy, with sildenafil and vardenafil possessing similar pharmacokinetic profiles and tadalafil exhibiting a longer duration of action. Phosphodiesterase-5 inhibitors should be avoided in patients using, or likely to use, nitrates because of potentiation of their hypotensive effect.

Intracavernosal injection of vasodilatory agents such as alprostadil and papaverine, and intraurethral alprostadil are useful second-line treatment options, local penile pain being the limiting side-effect of the latter. Self-injection can result in successful erections in patients without a vascular cause of impotence. The injection method needs to be adequately taught and should not be used more than once a week to reduce the risk of potential side-effects, such as corporeal fibrosis and priapism.

Variable results have been seen with vacuum devices, which act by increasing corporeal blood flow and are used in conjunction with a penile ring.

Penile prostheses, either semi-rigid or malleable rods inserted into the corpora, or self-inflatable prosthetic devices have been used with varying degrees of success. Mechanical failure, the need for manual inflation from an internal fluid reservoir and the length of waiting lists for the procedure in the UK have resulted in diminished use of this treatment option.

Further reading

Achermann JC, Ozisik G, Meeks JJ et al. Genetic causes of human reproductive disease. *J Clin Endocrinol Metab* 2002;87:2447–54.

Comprehensive review of genetic disorders that can lead to reproductive disease.

American Association of Clinical Endocrinologists. Medical guidelines for the evaluation and treatment of male sexual dysfunction: a couple's problem – 2003 update. *Endocr Pract* 2003;9:77–95. <https://www.aace.com/pub/pdf/guidelines/sexdysguid.pdf> [Accessed October 2013].

Well-structured approach to male sexual dysfunction.

Bhasin S, Cunningham GR, Hayes FJ et al. Task Force, Endocrine Society. Testosterone therapy in men with androgen deficiency syndromes: an endocrine society clinical practice guideline. *J Clin Endocrinol Metab* 2010;95:2536–59.

A useful evidence-based update on the guidelines for investigation and management of androgen deficiency syndromes.

Brinkmann A, Jenster G, Ris-Stalpers C et al. Molecular basis of androgen insensitivity. *Steroids* 1996;61:172–5.

Details of molecular physiology relevant to androgen insensitivity.

Diamanti-Kandarakis E, Bourguignant JP, Giudice LC et al. Endocrine-disrupting chemicals: an endocrine society scientific statement. *Endocr Rev* 2009;30:293–342.

A comprehensive review of endocrine disrupting chemicals and their effects on both male and female reproductive function. In terms of the male, this is based on animal studies and epidemiological data, and reviews the potential detrimental actions of dioxins, phthalates and numerous other disruptors on spermatogenesis, fetal urogenital development and testicular germ cell cancer.

Fazio L, Brock G. Erectile dysfunction: management update. *Can Med Assoc J* 2004;170:1429–37.

Explores current knowledge on the assessment and treatment of erectile dysfunction.

Kronenberg HM, Melmed S, Polonsky KS et al. editors. *Williams textbook of endocrinology*. 11th ed. Philadelphia: Saunders; 2007.

Includes a comprehensive review of male reproductive endocrinology.

Ostrer H. Sexual differentiation. *Semin Reprod Med* 2000;18:41–9.

Helps understanding of the factors at play during sexual differentiation.

Qaseem A, Snow V, Denberg TD et al. Hormonal testing and pharmacologic treatment of erectile dysfunction: a clinical practice guideline from the American College of Physicians. *Ann Intern Med* 2009;151:639–49.

An up-to-date review of the literature around investigation and treatment of ED with a sensible evidence-based discussion of the most appropriate investigation algorithms and current treatments.

Sultan C, Lumbroso S, Paris FI. Disorders of androgen action. *Semin Reprod Med* 2002;20:217–28.

This paper helps to further understand how disordered androgen action can manifest clinically.

Whitehead S, Miell J. *Clinical endocrinology*. Banbury: Scion; 2013.

A useful and easily read text focusing on basic science and clinical application. Useful chapter on disorders of gonadal function with a case based approach to investigation and management.

World Health Organization. Laboratory manual for the examination of human semen and semen-cervical mucus interaction. 5th ed 2010. <http://www.who.int/reproductivehealth/publications/infertility/9789241547789> [Accessed October 2013].

A useful volume that attempts to standardize tests and investigation in common use throughout the world.

APPENDIX 23.1: PROTOCOLS FOR ENDOCRINE INVESTIGATIONS

(i) Gonadotrophin releasing hormone stimulation test

In the investigation of gonadotrophin deficiency, assessment of the responses to exogenous GnRH may be useful. There is no need for the patient to fast unless the test is being combined with assessment of other anterior pituitary hormone responses to insulin-induced hypoglycaemia.

Gonadotrophin releasing hormone (100 µg) is given intravenously at time 0 and samples taken at 0, 20 and 60 min. Normally, LH and FSH concentrations both rise, though the degree of increase is dependent on the method used for measurement. The peak response may be seen at either 20 or 60 min.

The interpretation of the GnRH test is largely based on the basal values. In a patient with delayed puberty, a normal response or a response in which the LH peak is greater than that of FSH, suggests that the patient is

about to go into puberty. The response to GnRH may be suppressed by intervening illness.

Reference

Besser GM, Ross RJM. Are hypothalamic releasing hormones useful in the diagnosis of endocrine disorders? In: Edwards IR, Lincoln CRW, editors. *Recent advances in endocrinology and metabolism*. Edinburgh: Churchill Livingstone; 1989, pp. 135–158.

(ii) Clomifene test

The clomifene test may be helpful in distinguishing gonadotrophin deficiency from weight-related hypogonadism and idiopathic delayed puberty. Side-effects include visual disturbances, symptoms of oestrogen deficiency and depression, which may be severe enough to warrant discontinuation of the test.

Clomifene 50 mg is given daily for 5 days. Serum LH and FSH are measured on days 0 and 7; in addition, in females, progesterone concentration is measured on day 21 to confirm ovulation. In a normal response, LH and FSH concentrations rise to outside the reference ranges or to double the basal values. A lack of response suggests gonadotrophin deficiency due to pituitary or hypothalamic disease. Patients with anorexia nervosa may show no response. Prepubertal children show no response and children in early puberty may actually show a reduction in gonadotrophin concentrations during the test.

Reference

Andersen DC, Marshall JC, Young JL et al. Stimulation tests of pituitary–Leydig cell function in normal male subjects and hypogonadal men. *Clinical Endocrinology* 1972; 1:127–40.

(iii) Human chorionic gonadotrophin stimulation test

In the differential diagnosis of male hypogonadism, a stimulation test with hCG may be useful. There are no specific precautions and the test does not have to be carried out after fasting.

Human chorionic gonadotrophin is injected i.m. in a dose of 2000 IU on days 0 and 2 and serum testosterone measured on days 0, 2 and 4. A normal response is defined as a rise in serum testosterone to within the reference range. A failure of any rise in testosterone suggests an absence of functioning testicular tissue. If the testes are not palpable in the scrotum or inguinal canals, a rise in testosterone suggests the presence of intra-abdominal testes. In gonadotrophin deficiency with normal testes, the low basal testosterone concentration should triple after hCG.

Reference

Andersen DC, Marshall JC, Young JL et al. Stimulation tests of pituitary–Leydig cell function in normal male subjects and hypogonadal men. *Clinical Endocrinology* 1972; 1:127–40.

Inherited metabolic disease

Fiona Carragher • Mike Champion

CHAPTER OUTLINE

INTRODUCTION 461

CLINICAL PRESENTATION AND PATHOPHYSIOLOGY 461

- Neonatal presentation 462
- Presentation at weaning 464
- Presentation in later infancy 464
- Presentation at puberty 464
- Presentation during adulthood 465
- Presentation during pregnancy 465
- Presentation postpartum 466

NEWBORN SCREENING 466

INHERITANCE 466

- Autosomal recessive inheritance 466
- Autosomal dominant inheritance 467
- X-linked inheritance 467
- Mitochondrial inheritance 467

DIAGNOSTIC STRATEGIES 468

- Essential laboratory investigations 469
- Second-line investigations 472

Functional and loading tests 474

Confirmatory investigations 475

PRENATAL DIAGNOSIS 476

MANAGEMENT 476

- Strategies to replace a missing product 476
- Inhibition of product breakdown 478
- Enzyme replacement therapy 478
- Organ transplantation 479
- Gene therapy 481
- Other molecular therapies 481
- Strategies to reduce the formation of toxic metabolites 481
- Blockage of site of action of toxic metabolites 482
- Strategies to remove toxic substances 482
- Additional treatments 483
- Substrate depletion 483
- Substrate deprivation 483

CONCLUSION 483

INTRODUCTION

Inherited metabolic diseases (IMDs), also known as in-born errors of metabolism, are inherited conditions that develop as a result of mutations that affect the function of proteins. The majority of IMDs are monogenic conditions and the mutant proteins are enzymes, but others involve structural proteins, receptors, hormones or transport proteins. Although they are inherited, not all IMDs present in the newborn period: some present later in childhood or not until adult life.

Inherited metabolic diseases are individually rare. Of those that present in childhood, the commonest, such as phenylketonuria (PKU) and medium chain acyl-CoA dehydrogenase deficiency (MCADD), have an incidence of 1 in 10 000. However, collectively, the IMDs that present in childhood are thought to occur with an incidence of 1 in 750 live births, although the true incidence is unknown, as no newborn screening programme is comprehensive, and many of these conditions go undiagnosed. More diagnoses are being secured with advancing diagnostic techniques, such as tandem mass spectrometry, and the growing awareness of these disorders by clinicians. Similarly, the therapeutic options are continuing

to expand, increasing the pressure to detect cases at an earlier stage. Importantly, improving treatments leads to longer survival that, in turn, may present new clinical challenges such as the management of pregnancy in an affected mother. Altering the natural history of a condition may also reveal previously unknown long-term complications.

If IMDs that present in adult life (e.g. familial hypercholesterolaemia, genetic haemochromatosis) are included, at least 1 in 100 individuals has one of these conditions. If disorders such as the haemoglobinopathies are included (see Chapter 28), the prevalence is greater still.

CLINICAL PRESENTATION AND PATHOPHYSIOLOGY

Inherited metabolic diseases may present at any age. However, there are key times when presentation is more common, either owing to the individual having to survive biochemically without the support of the mother's placenta, or to additional metabolic stresses at that particular time.

Neonatal presentation

Many IMDs present in the neonatal period. They may be considered in four broad categories: problems of synthesis and breakdown of complex molecules; intoxications; energy deficiency states, and seizure disorders.

Defects in synthesis and breakdown

Many complex molecules are integral to cell-to-cell communication and ordered patterning within the developing embryo. Failure to make these complex molecules can, therefore, result in disordered embryogenesis, presenting as a dysmorphic neonate at birth. An example is Zellweger syndrome, the most severe of the peroxisomal biogenesis defects. Affected neonates have a typical appearance, with a large fontanelle, prominent forehead and hypertelorism, hypotonia, hepatomegaly and calcific stippling, particularly of the knees and shoulders, on X-rays. The principle defect in peroxisomal disorders lies within the *PEX* genes, which encode the peroxin proteins critical for the targeting and importing of peroxisomal enzymes and proteins into peroxisomes, resulting in multiple enzyme deficiencies. This group of disorders is diagnosed by the analysis of very long chain fatty acids (VLCFAs), which are elevated in plasma owing to the block in their oxidation, which is a peroxisomal process. Management remains supportive, rather than curative.

Smith–Lemli–Opitz syndrome is another example of a synthetic defect. It results from a block in the penultimate step in cholesterol synthesis, 3β -hydroxysterol- $\Delta 7$ -reductase. The production of cholesterol, an essential component of cell membranes, is decreased with marked elevation of its precursor 7-dehydrocholesterol (7-DHC) in body fluids and tissues. The characteristic dysmorphism includes anteverted nares, low-set ears, micrognathia, ptosis and microcephaly. Other features include syndactyly of the 2nd and 3rd toes, present in 98% of patients, genital and renal anomalies, learning difficulties and severe failure to thrive. Dietary cholesterol replacement and statin treatment have been used with the aim of inhibiting cholesterol synthesis and reducing accumulation of 7-DHC; no convincing effects have been seen with either approach, but this may be because the phenotype is determined by the in utero availability of cholesterol.

Problems with the breakdown of complex molecules result in storage disorders. These tend not to be apparent at birth, but rather become so with time as the substance(s) stored in excess begins to affect structure and function. For example, affected children with Hurler syndrome (mucopolysaccharidosis type I, MPSI) appear normal at birth, but the gradual accumulation of glycosaminoglycans over time produces the typical coarsening of the features, corneal clouding, organomegaly and dysostosis multiplex (distortion of the normal bony architecture secondary to storage material) that are characteristic of the condition. However, it is parental concerns about delayed development, rather than the coarse features, that usually bring these patients to medical attention. Storage disorders in which dysmorphism and organomegaly are

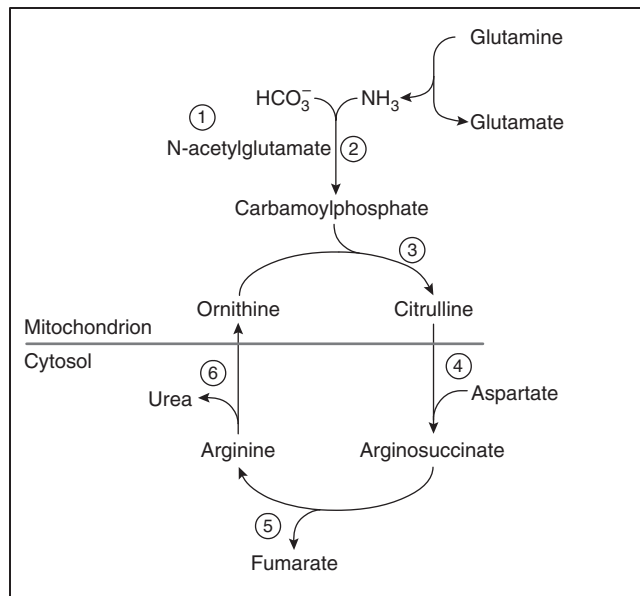
often present in the first month of life include I-cell disease, infantile sialic acid storage disease and early infantile GM1 gangliosidosis.

Intoxications

Intoxication is the classic presentation of inherited metabolic disorders. An unremarkable period immediately after birth is followed by increasing clinical abnormality as the baby feeds, and toxic metabolites, which cannot be broken down because of the metabolic block, accumulate. Prior to birth, these toxic metabolites are cleared via the placenta. Typically, clinical features develop within 48–72 h after birth, but can take considerably longer. Phenylketonuria (phenylalanine hydroxylase deficiency) is an intoxication, but the first signs of the condition do not usually develop until 6–12 months after birth when motor developmental milestones are not met. Phenylketonuria (PKU) also illustrates two other features of inherited metabolic diseases affecting enzymes: accumulation of the substrate of the defective enzyme may lead to increased metabolism via alternative, normally minor, pathways by a mass action effect (phenylalanine is transaminated to phenylpyruvate and phenylketones), and there may be deficiency of the normal product of the enzyme (in this case, tyrosine). Both of these may contribute to the clinical presentation.

Urea cycle defects typically present with encephalopathy in the first days of life. As milk feeds are established, the block in the conversion of waste nitrogen, derived from the amine groups of amino acids, to urea produces hyperammonaemia and increases glutamine formation. Ammonia interferes with neurotransmission causing astrocyte swelling, and glutamine increases the risk of cerebral oedema, owing to the osmotic load as it accumulates in the brain. Ammonia is also a respiratory stimulant, acting on the respiratory centre in the brain stem to produce a respiratory alkalosis, which is an unusual finding in a sick neonate. Diagnosis relies on measuring plasma ammonia concentration in conjunction with plasma amino acids and urinary orotic acid to determine the location of the block (Fig. 24.1). Immediate management depends on attempting to increase the clearance of ammonia and supplementation of arginine, usually a non-essential amino acid, synthesized in the urea cycle, which becomes an essential amino acid in many of the urea cycle defects. Low arginine also contributes to neurotoxicity by reducing nitric oxide and creatine synthesis.

Galactosaemia (galactose 1-phosphate uridylyltransferase deficiency) usually presents a little later, at the end of the first week of life, with jaundice due to conjugated bilirubinaemia, hepatomegaly, coagulopathy and characteristic ‘oil drop’ cataracts. The diagnosis is made by measuring the enzyme activity in red cells. The exact pathogenesis is not fully elucidated, but restriction of galactose and lactose (a disaccharide of glucose and galactose) is effective in reversing the liver toxicity. However, the diet does not prevent all adverse effects of the disease and it is clear that endogenous galactose production is significant. As with many IMDs, some infants present



Enzyme	Disorder	Amino acids	Orotic acid
1 N-acetylglutamate synthetase (NAGS)	NAGS deficiency	↑glutamine	Normal
2 Carbamoylphosphate synthetase (CPS)	CPS deficiency	↑glutamine ↓arginine	Normal
3 Ornithine transcarbamoylase (OTC)	OTC deficiency	↑glutamine ↓arginine	↑↑
4 Argininosuccinate synthetase	Citrullinaemia	↑↑citrulline ↓arginine	↑
5 Argininosuccinate lyase	Argininosuccinic aciduria	↑↑citrulline ↑↑argininosuccinic acid ↓arginine	↑
6 Arginase	Hyperargininaemia	↑arginine	↑

FIGURE 24.1 ■ The urea cycle. The numbers refer to the enzymes listed in the table, deficiencies of which cause changes in the plasma concentrations of amino acids and increased urinary excretion of orotic acid as indicated.

later than the neonatal period, with the renal effects of galactosaemia, namely proximal tubulopathy and rickets, rather than the hepatic consequences. A similar difference in presentation is seen in tyrosinaemia type 1, with an early hepatotoxic picture not unlike galactosaemia and a later Fanconi-type presentation due to nephrotoxicity. The key diagnostic marker in tyrosinaemia type 1 is the detection of succinylacetone on urinary organic acid analysis.

Energy deficiency disorders

Energy deficiency disorders are the result of either a fundamental block in energy production, as seen in the congenital lactic acidoses, or a failure of adequate energy production in the absence of a regular food supply. The congenital lactic acidoses have a number of causes (Box 24.1), with a definitive diagnosis being made in only approximately half of the patients. Clinical features develop early, but significant problems due to metabolic decompensation may take longer to develop, in comparison with intoxications.

BOX 24.1 Some inherited metabolic disorders causing lactic acidosis

Primary lactic acidosis

- Respiratory chain disorders
- Pyruvate dehydrogenase deficiency
- Disorders of gluconeogenesis (e.g. fructose 1,6-bisphosphatase deficiency)

Secondary lactic acidosis

- Organic acidurias
- Urea cycle disorders
- Fatty acid oxidation disorders
- Glycogen storage disease type I

The key diagnostic marker is a raised plasma lactate concentration but, in practice, secondary hyperlactataemia (due, e.g. to hypoxia, hypovolaemia, hypotension etc.) is more common. The distinguishing diagnostic clue is the absence of ketones in the secondary elevations. Spurious elevation may result from squeezing the arm to elevate the vein prior to venepuncture. A free-flowing sample is required and an arterial stab should be used if a free-flowing sample cannot be obtained otherwise. Further evidence may be gleaned from measuring lactate in the cerebrospinal fluid (CSF), which reflects brain lactate over days rather than at that specific time. Secondary elevations of CSF lactate are seen in primary brain infections such as encephalitis and meningitis, and following seizures.

In certain disorders, energy deficiency only becomes apparent when feeding is interrupted. Neonates with fatty acid oxidation defects remain asymptomatic while fed, but can develop hypoketotic hypoglycaemia during prolonged fasting or intercurrent infections. The commonest fat oxidation defect, MCADD, usually presents much later, at around the age of one year, but a subgroup present in the first few days of life, before feeding becomes established, most commonly if they are breast fed. Neonatal glycogen stores are more rapidly exhausted than those in older infants and adults, requiring mobilization of fat stores for energy. Fats cannot be directly utilized by the brain, requiring conversion in the liver to ketones, which can be. Failure to produce ketones results in encephalopathy as a result of a combination of hypoglycaemia and the accumulation of acylcarnitines. Clinically evident hypoglycaemia is a late feature: treatment, either with oral glucose polymer or intravenous 10% dextrose, should be started before it occurs.

Seizure disorders

Inherited metabolic diseases are a rare cause of neonatal seizures compared with birth asphyxia and infections, and are usually a late, non-specific feature of blocks in intermediary metabolism. There are, however, a number of IMDs that typically present with seizures at this time that will not be detected on routine investigation and so need to be specifically excluded if no other cause is apparent (Table 24.1). Seizures may have been present in the antenatal period and been interpreted as the

TABLE 24.1 Inherited metabolic disorders presenting with neonatal seizures and their diagnostic investigations

Disorder	Investigation
Biotinidase deficiency	Plasma biotinidase Urine organic acids CSF:plasma glycine ratio
Non-ketotic hyperglycinaemia (NKH)	
3-Phosphoglycerate dehydrogenase deficiency	CSF serine
Pyridoxine 5' phosphate oxidase deficiency	CSF neurotransmitters Urine organic acids Trial of pyridoxal phosphate PNPO genotyping
Pyridoxine-dependent seizures	Trial of pyridoxine Plasma α -amino adipic semialdehyde Antiquitin genotyping
Purine disorders	Urinary purine studies
Sulfite oxidase/molybdenum cofactor deficiency	Dip-stick test for sulphite on fresh urine Urinary purine studies
Peroxisomal disorders	Plasma very long chain fatty acids

baby being particularly active because of increased fetal movements. Some mothers do note the rhythmic nature of these movements.

Presentation at weaning

Weaning is a time when new dietary components may first be encountered, and if any one of their metabolic pathways is blocked, clinical features may develop. Weaning may also result in greater intake of the particular substrate whose metabolism is compromised, e.g. protein. If the relevant metabolic pathway is unable to cope with the increased load, the threshold for the development of clinical features may be breached. An example of the latter is a partial urea cycle defect. Prior to weaning, protein intake may have been within the tolerance of the compromised pathway, but with the addition of solids, greater quantities of protein can be ingested. If, as a result, the pathway's capacity is exceeded, hyperammonaemia will ensue. A similar effect is seen in children with hyperphenylalaninaemia detected on newborn screening. A decision is usually made to monitor concentrations to see if the phenylalanine increases with greater food intake, particularly around the time of weaning.

Hereditary fructose intolerance (aldolase B deficiency) does not present until an infant is exposed to fructose, which typically occurs around the time of weaning when pureed fruit is introduced; the sucrose in fruit is broken down to fructose and glucose. There is no fructose or sucrose in breast milk or formula milk. In the affected patient, fructose 1-phosphate accumulates, inhibiting glucose production, promoting hypoglycaemia and depleting inorganic phosphate, thereby reducing ATP production. Clinically, the first features are nausea and vomiting with postprandial

hypoglycaemia. If the condition is not recognized, and fructose ingestion persists, the infant will fail to thrive and may develop liver and kidney failure. Confirmatory enzymology requires liver biopsy; however, in some patients a clinical diagnosis can be made on the basis of a history of exposure to fructose and reversal of symptoms after elimination of sources of fructose from the diet. The diagnosis may then be confirmed by genotyping.

Presentation in later infancy

Infancy is a time when there is a considerable risk of infections, as the body is exposed to numerous infectious agents for the first time and the immune system is still developing. Infection causes increased metabolic stress; patients with known IMDs may decompensate and others may present for the first time. Some IMDs show seasonal variation in presentation: for example, MCADD presents more frequently in the autumn and winter months, because of the increase in infections at this time of year.

Glutaric aciduria type I (GA-I) is an autosomal recessive defect in lysine catabolism. Affected infants may have large heads but minimal neurological signs prior to catastrophic metabolic decompensation around the end of the first year, usually precipitated by an infection. Such decompensation causes damage to the basal ganglia, with resultant irreversible dystonia and movement disorder. In presymptomatic siblings, prospective early aggressive management of intercurrent illnesses with antibiotics, protein (lysine) restriction, carnitine supplements and hyperalimination with glucose polymer or intravenous dextrose, can reduce the incidence of decompensation and subsequent neurological sequelae. Measurement of urinary organic acids shows elevated glutarate and 3-hydroxyglutarate in this condition. Occasional patients have been described with classic histories and abnormal enzymology in fibroblast studies without the typical urine abnormalities; plasma acylcarnitine analysis may show reduced free carnitine and a glutaryl carnitine peak, but may be normal in such patients.

Infants may present with IMDs towards the end of the first year, when growth slows. On average, a baby will put on just under 7 kg in weight in the first year of life, compared with 2 kg each subsequent year during childhood. This means that for the same protein intake, more will need to be catabolized, as less is needed for growth. This increased pressure on an affected metabolic pathway may result in decompensation.

Presentation at puberty

Puberty is recognized as a difficult time for teenagers and rebellion affects all areas of life, including caring for their health. Patients with significant IMDs may break their diets or fail to take their medication in an attempt to be 'normal'. Both can precipitate decompensation. However, new presentations are also seen at this time, probably as a result of changes in growth and the hormonal milieu. For girls with urea cycle defects, it is well recognized that following menarche, symptoms may

fluctuate in time with the menstrual cycle, being worse in the few days leading up to, and including, the start of the period. The use of hormonal therapy to suppress ovulation and menstruation has proven helpful in some patients.

Presentation during adulthood

Inherited metabolic diseases are often considered to be paediatric conditions, but it must be remembered that they can present at any age. This may be owing to the defect being less severe so that an individual's metabolism has not previously been sufficiently stressed to provoke decompensation. Patients with partial ornithine transcarbamoylase (OTC) deficiency, the commonest of the urea cycle defects, may remain asymptomatic throughout childhood and only present in adult life. Some adults may not have been exposed previously to sufficient quantities of metabolites they cannot deal with. Some adults with hereditary fructose intolerance will have learned at an early age that sweet, sugary foods make them feel unwell and will subconsciously avoid them. Many patients with this condition have perfect dentition owing to their self-imposed diet.

For some IMDs, presentation differs between adults and children. The severe form of X-linked adrenoleukodystrophy (ALD) usually presents at 5–10 years with progressive neurological deterioration, ultimately leading to spastic quadriplegia, seizures, vegetative state and death. More than 90% of affected boys have adrenal insufficiency. However, adrenal involvement may precede, or follow, neurological symptoms by years. There is wide phenotypic variability within families, and another family member may not present until adulthood. The cerebral form in adults is extremely rare, accounting for <5% of all cases. The commoner adult presentation is adrenomyeloneuropathy (AMN), with a slowly progressive neurological picture mimicking spinal cord syndrome, with spastic gait and difficulty voiding urine. Some 10% of patients have only adrenal involvement, and 10% remain asymptomatic. Adrenoleukodystrophy is the commonest peroxisomal disorder. The defective adrenoleukodystrophy protein (ALDP) is thought to play a role in the uptake of VLCFAs across the peroxisomal membrane, but the exact pathophysiology is not completely understood. The use of Lorenzo's oil (a 4:1 mixture of glyceryl trioleate and glyceryl trierucate) reduces the accumulation of the VLCFAs, but fails to prevent neurological decline in symptomatic patients. It appears to have a role in asymptomatic boys reducing the risk of developing abnormalities apparent on magnetic resonance imaging. Bone marrow transplantation can be performed to stabilize cerebral ALD in the very early stages of demyelination. The varied phenotype is not explained by differences in genotype, as family members with different manifestations of the condition will usually have the same mutation.

Adrenoleukodystrophy also demonstrates another variation in adult presentation, that of the manifesting heterozygote. This is seen in some X-linked conditions where a carrier female develops symptoms. Two-thirds of females carrying an ALD gene mutation have some

degree of neurological involvement, ranging from brisk reflexes and mild abnormalities on clinical examination to a full-blown AMN-like picture. This may be misdiagnosed as multiple sclerosis.

A similar phenomenon is seen in Fabry disease (α -galactosidase deficiency). Deposition of glycosphingolipids in blood vessel walls, heart, kidneys, skin and autonomic ganglia produces cerebrovascular disease, cardiac disease, nephropathy, angiokeratoma, corneal dystrophy and acroparaesthesia (severe pain in the extremities). It was believed that the majority of female carriers were entirely asymptomatic throughout life, but it is now clear that one-third of female carriers have significant symptoms, which may be so severe as to warrant treatment. One explanation for the spectrum of severity is the random inactivation of the X chromosome in cells – lyonization. If sufficient numbers of the unaffected X chromosomes are suppressed, a significant number of cells will have a functioning, affected X chromosome. The resulting enzyme function may be low enough for symptoms to develop.

A significant number of IMDs present for the first time in adulthood. The classic mitochondrial syndromes that led to the recognition of mitochondrial DNA (mtDNA) mutations and their role in pathology are all primarily adult presentations: an example is Leber hereditary optic neuroretinopathy, presenting most commonly in the third decade of life with bilateral, painless central vision loss.

The IMDs that typically present only in adult life usually involve the accumulation of a toxic substance that, although it begins at birth, takes many years to become manifest. Important examples include heterozygous familial hypercholesterolaemia (homozygotes usually present with xanthomata or coronary disease in the second decade), familial combined hyperlipidaemia and primary (genetic) haemochromatosis (excepting the rare neonatal variant). All these conditions are discussed elsewhere in this book.

Presentation during pregnancy

The physiological stresses of pregnancy may precipitate crises in women with IMDs so meticulous care is required to ensure the best outcome for the mother and baby. Some women develop symptoms in pregnancy because they carry an affected child, although they do not have the condition themselves. The classic examples of this are the long chain fatty acid oxidation defects, long chain hydroxyl-acyl-CoA dehydrogenase deficiency (LCHADD) and very long chain acyl-CoA dehydrogenase deficiency (VLCADD). These conditions are recessively inherited and the mother, therefore, is an obligate carrier with 50% enzyme activity, which is compatible with normal life and function. Clinical manifestation of most IMDs only occurs if activity is <5%. However, the mother also has to combat the metabolic stresses of pregnancy, which increase the load on the pathway. If the fetus is affected, she will also have to break down the long chain acylcarnitines that the fetus is producing owing to the block in its metabolism. Such women do not present with the typical hypoketotic hypoglycaemia seen in

affected infants, but liver function is impaired, precipitating HELLP (hepatomegaly, elevated liver enzymes and low platelets) syndrome or the rarer, more severe acute fatty liver of pregnancy (AFLP). This may require intensive care and early delivery of the fetus. Only a small number of women with HELLP and AFLP carry affected fetuses, but it is imperative that their newborn infants are screened by measurement of plasma acylcarnitines to exclude a fat oxidation defect. Screening only for the associated dicarboxylic aciduria, using urinary organic acid analysis, is not sensitive enough and has led to the diagnosis being overlooked.

Presentation postpartum

Women with partial OTC deficiency may present for the first time a few days after delivery. They are able to withstand the increased metabolic stress of pregnancy and delivery, but the massive protein load presented by the involution of the uterus precipitates hyperammonaemia. Many of these women have never had previous symptoms that might alert their obstetrician to the cause of their illness, although some have an aversion to high protein diets.

Clinical abnormalities in the infant may also reveal an undetected IMD in the mother. Maternal PKU syndrome is a description of the clinical consequences to the fetus of in utero exposure to elevated plasma phenylalanine resulting from PKU in the mother. A mother not known to have PKU, usually because she was born in an area without a neonatal screening programme, may have had hyperphenylalaninaemia that was sufficiently mild for her not to have come to medical attention, but significant enough to affect the fetus. Typical features are low birth weight, microcephaly, cardiac abnormalities and developmental delay. The infant does not have PKU or hyperphenylalaninaemia, and therefore does not require dietary restriction. The diagnosis is usually made when the infant is investigated for microcephaly or developmental delay. Plasma phenylalanine needs to be measured in the mother for diagnosis; urinary amino acids may not be sufficiently sensitive and diagnoses have been missed when only this investigation has been used.

NEWBORN SCREENING

Newborn screening is used to detect conditions that have a presymptomatic period during which treatment can dramatically improve outcome. In the UK, conditions screened for at this time include congenital hypothyroidism (see below), PKU, MCADD, cystic fibrosis and sickle cell disorders. A pilot study is underway in the UK to examine an expanded neonatal screening programme including maple syrup urine disease, glutaric aciduria type 1, isovaleric acidaemia, LCHADD and pyridoxine unresponsive homocystinuria. This is modest compared to some countries, e.g. the USA where newborns are screened for over 30 conditions. However, for many of these conditions the natural history is not fully understood, treatments are only partially effective and

infants may present and die prior to the screening result being available. The detection of infants who will never develop symptomatic disease remains a real concern. During the development of MCADD screening, mutations were identified in infants with raised octanoylcarnitine that have never been associated with clinical abnormalities.

Screening is also used to detect IMDs in populations with higher frequencies of a particular condition, usually secondary to a founder effect (a mutation occurring early in the settlement of a geographically isolated area or within a limited gene pool so that it occurs with high frequency), for example Tay–Sachs disease in the Ashkenazi Jewish population.

Newborn screening is widely practised for congenital hypothyroidism, but while some of these infants have inherited disorders of thyroid hormone synthesis, some have thyroidal agenesis or dysgenesis (that is, failure of development of the gland), for which a genetic basis has not been defined.

INHERITANCE

The molecular bases of IMDs are mutations in genes that adversely affect the functions of specific proteins. The patterns of inheritance vary, but the majority of these conditions are autosomal recessive.

Autosomal recessive inheritance

Autosomal recessive inheritance requires both parents to carry a mutation affecting the same gene. It is estimated that we each carry 250–300 loss of function mutations in our genes, but, as we also have a normal copy of the gene, the resultant 50% activity is more than sufficient for normal function: as a result, carriers of autosomal recessive conditions are generally not clinically affected (patients with IMDs typically have <5% activity). Even if the parents are carriers for the same condition, both faulty copies of the gene need to be passed on to the embryo, resulting in a 1 in 4 risk of the infant being affected in each pregnancy (Fig. 24.2A). There is also a 1 in 4 chance that neither faulty gene will be inherited and a 2 in 4 (1 in 2) chance that the embryo will be a carrier.

If the frequency of the gene defect in the general population is known, the incidence can be calculated: for example the carrier frequency for PKU is 1 in 50, so the incidence equals $(1/50 \times 1/50)$ (the chance of two carriers having children together) multiplied by 1/4 (the chance of them having an affected child), that is, 1/10000. If a certain condition is known to exist in a family, genetic counsellors can use calculations of this sort to inform individual couples of their risk of having an affected child. For example, if the sister of an individual PKU were to have a child with an unrelated man, the risk of the baby having the condition would be 1 in 300, which is considered negligible. (She is not affected: she has a 2 in 3 chance of being a carrier and 1 in 3 of being homozygous normal – the denominator is three rather than four because she is unaffected). The man's chance of being

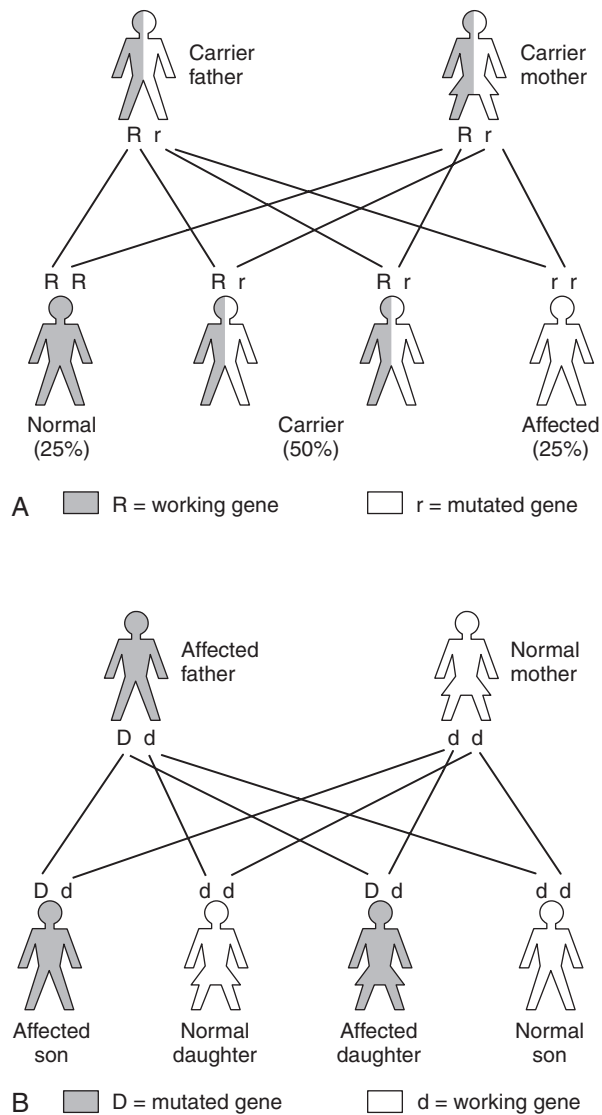


FIGURE 24.2 ■ (A) Autosomal recessive inheritance. (B) Autosomal dominant inheritance.

a carrier is 1 in 50 (the carrier frequency in the general population), so the risk of two faulty genes being passed to an embryo is $\frac{2}{3} \times \frac{1}{50} \times \frac{1}{4} = \frac{1}{300}$.)

Consanguinity within a family increases the risk of autosomal recessive conditions and there are some disorders with a high frequency within particular populations. There may be a founder effect within certain populations, for example the Pennsylvania Mennonites, in whom the severe form of maple syrup urine disease has an incidence of 1:176. In these kindred, lineage can be traced to one couple who emigrated from Europe in the eighteenth century. The tradition of first cousin marriages in some immigrant groups is increasing the incidence of IMDs in some areas of the UK.

It should be noted that, although affected individuals with autosomal recessive IMDs are often classified as homozygotes, the existence of multiple mutations affecting individual genes means that some are, strictly speaking, compound heterozygotes.

Autosomal dominant inheritance

Autosomal dominant inheritance is rare in IMDs with the exception of some of the porphyrias, for example acute intermittent porphyria, hereditary coproporphyria and porphyria variegata (see Chapter 28). The risk of transmission in autosomal dominant disorders is 1 in 2, as only one faulty copy of the gene is required to cause disease (Fig. 24.2B), although variable penetrance and other factors may influence the degree to which the offspring are affected. Dominant inheritance is less common than recessive, as severe disease may result in death prior to an individual reaching reproductive maturity and so the defective gene fails to pass on to the next generation. However, in some cases, the mutation may have occurred *de novo* in the embryo. For some dominantly inherited conditions, the condition may provide some protection against another serious condition (e.g. the sickle haemoglobin trait provides protection against *P. falciparum* malaria, see Chapter 29).

X-linked inheritance

X-linked recessive inheritance occurs with a variety of IMDs, for example: OTC deficiency; pyruvate dehydrogenase complex deficiency; Hunter syndrome (mucopolysaccharidosis type II); Lesch–Nyhan syndrome (a purine disorder); Fabry disease (sphingolipidosis), and ALD. This pattern of inheritance is characterized by carrier females (unaffected, as they have a normal copy of the gene on their other X chromosome) passing the gene to their affected sons (the Y chromosome does not carry the gene so there is no normal copy) (Fig. 24.3A). The chance of a son being affected is 1 in 2 in each pregnancy, the same as the risk of a daughter being a carrier. As the affected gene lies on the X chromosome, affected fathers can only pass it to their daughters, who will be obligate carriers.

Unlike autosomal recessive conditions, in X-linked disorders carriers may manifest the disorder clinically, for example in OTC deficiency. As discussed previously, in carrier females the variation in the degree of symptoms is due to lyonization, the random inactivation of one of the two X chromosomes in all cells, including hepatocytes. The severity of disease depends on the percentage of hepatocytes expressing the normal gene. This can lead to a varied clinical presentation within families, with some female carriers presenting with severe hyperammonaemia in the neonatal period and others being apparently unaffected.

X-linked dominant conditions require only one copy of the gene to be inherited to express the condition, hence males and females are equally affected, for example vitamin D-resistant rickets (Fig. 24.3B).

Mitochondrial inheritance

Mitochondria are unique intracellular organelles, in that they have their own genes. However, a fully functional mitochondrion is the product of both the nuclear and mitochondrial genomes with the vast majority of genes, in the order of 1300, being encoded in the nucleus.

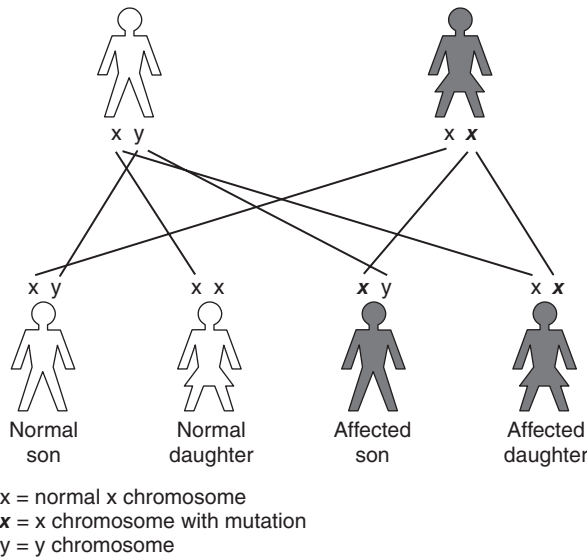
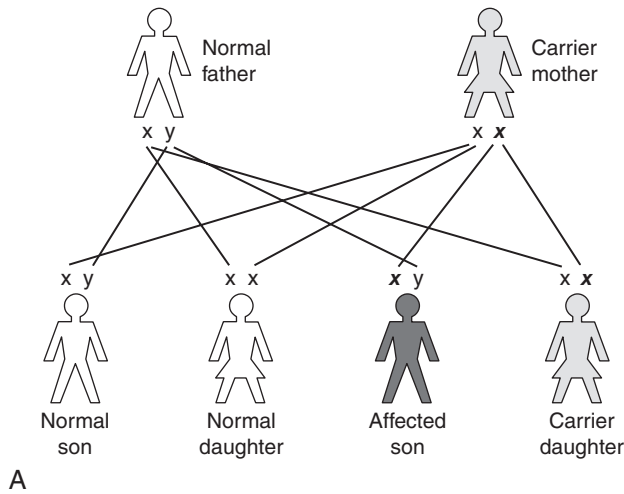


FIGURE 24.3 ■ (A) X-linked recessive inheritance. (B) X-linked dominant inheritance.

If disease-causing mutations arise in the nucleus, these can be inherited in the usual way, for example as autosomal recessive or dominant, or X-linked traits. Mitochondrial DNA (mtDNA) is inherited in a matrilineal fashion; that is, mtDNA is inherited exclusively from the mother. Paternal mitochondrial DNA is present in the mitochondria in the sperm's tail to provide ATP for propulsion. On fertilization, the tail of the sperm is left outside and hence no paternal mtDNA enters the zygote. Mitochondrial DNA mutations can therefore be inherited only from females, but can affect males or females. Point mutations inherited in this fashion include those for: myoclonic epilepsy, ragged red fibres (MERRF); mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes (MELAS) and Leber hereditary optic neuropathy.

The degree to which the offspring will be affected in these conditions is influenced by the mutant load, in that each cell has many mitochondria and each mitochondrion has multiple copies of mtDNA, a mix of

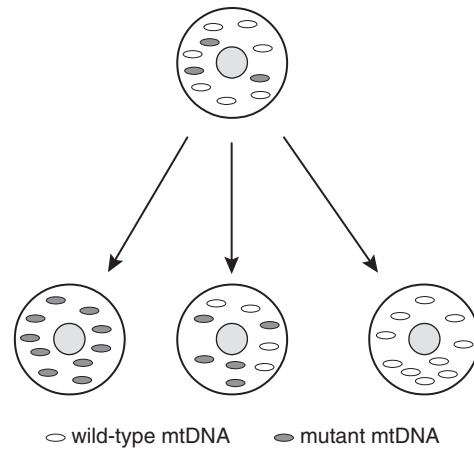


FIGURE 24.4 ■ Mitochondrial DNA replication. Mitochondria randomly segregate to daughter cells following replication. Each cell may contain mitochondria with both mutant and wild-type DNA (heteroplasmy). Daughter cells may drift towards homoplasmy for either wild-type or mutant mtDNA. As the percentage mutant mtDNA load increases, the chance of developing clinical features increases.

normal (wild type) and mutant. The presence of normal and mutant mtDNA within the same cell is called heteroplasmy. During cell division, the mtDNA becomes randomly divided between the daughter cells (Fig. 24.4). The proportion of mutant mtDNA may drift towards homoplasmy of either normal or mutant mtDNA. When the proportion of mutant mtDNA reaches a certain threshold within an individual cell, cellular function is compromised and, if sufficient cells are similarly affected, clinical features will develop. The distribution of the affected mitochondria within the body will determine the presentation. Mitochondrial disorders typically present with multiorgan involvement; however, the high energy-requiring organs are more frequently involved, especially brain, muscle, liver, heart, kidneys and eyes. Cells that divide rapidly may clear the mutant mtDNA because cells with normal function have reproductive advantage, e.g. refractory anaemias and enteropathy can resolve as the marrow or gastrointestinal tract recover, whereas brain and muscle involvement tend to be progressive.

Rearrangements (deletions/duplications) within the mitochondrial genome are usually sporadic and therefore the risk of recurrence is small. Examples include Kearns–Sayre syndrome (external ophthalmoplegia, pigmentary retinopathy, heart block) and Pearson syndrome (anaemia and pancreatic dysfunction, both endocrine and exocrine).

DIAGNOSTIC STRATEGIES

Close collaboration between the clinician and the metabolic laboratory is essential to establish the diagnosis of an IMD promptly. Links should be established between referring centres and specialist metabolic units, so that urgent cases can be discussed, investigations planned and a structured approach taken to the use of small samples.

Details of the patient's drug and feeding history, as well as transfusion status, will be needed for a full interpretation of results. Ideally, in the specialist unit, the clinical and laboratory services should be integrated with a multidisciplinary approach to the investigation and interpretation of results. This should lead to rapid diagnosis and initiation of appropriate treatment.

Essential laboratory investigations

When the suspicion of an IMD is raised, there are a number of basic laboratory investigations that can be performed immediately in the local hospital. These analyses will include full blood count and serum electrolytes as well as the more specific investigations discussed below. Their results may provide clues to the diagnosis and suggest further investigations.

Blood gas analysis

Blood gas analysis is a key investigation in a patient with a suspected IMD, particularly during acute episodes of illness. The most common acid–base imbalance seen in the paediatric population is a metabolic acidosis, secondary to infections, catabolic states or severe dehydration. The difficulty in interpretation in these circumstances is that these conditions are often the triggers of acute decompensation in patients with IMD. Calculation of the anion gap may assist in interpretation of the blood gas analysis: metabolic acidosis with an increased anion gap (>16 mmol/L) is observed in many IMDs, whereas metabolic acidosis with a normal anion gap is more likely to be due to diarrhoea or renal tubular acidosis.

In patients with a metabolic acidosis and suspected IMD it is important to test the urine for ketones. Point-of-care testing for urine and blood ketones, using simple stick technology, is readily available in most neonatal and paediatric units and provides a rapid indication of the ketotic status of the patient. Ketosis is a physiological response to fasting, and is often seen in normoglycaemic children with severe nausea and vomiting. However, during the neonatal period, metabolic acidosis with ketonuria is nearly always pathological and should prompt consideration of branched chain amino acid disorders: propionic acidaemia, methylmalonic acidaemia, isovaleric acidaemia and maple syrup urine disease, as possible diagnoses. Urinary organic acids, blood acylcarnitine and amino acid analysis should be undertaken urgently to elucidate the underlying defect.

The detection of urine ketones, in addition to blood lactate and glucose measurement, provides the basis for the differential diagnosis of metabolic acidosis in the context of IMD (Box 24.2). Although diabetes mellitus is not an IMD, it is the most important diagnosis to consider when ketoacidosis is found associated with hyperglycaemia. In ketotic patients with hypoglycaemia, the gluconeogenic and glycolytic defects should be considered first. The most suggestive features in this group are hepatomegaly and lactic acidaemia, although neither is consistently present. If the ketosis is sustained, the rare ketolysis defects (succinyl-CoA:3-oxoacid-CoA

BOX 24.2 Differential diagnosis of metabolic acidosis

Acquired causes

- Circulatory failure (cardiac abnormalities, hypovolaemia, sepsis)
- Renal failure
- Hypoxaemia

Inherited causes

- Organic acidaemias
- Lactic acidaemias (see Box 24.1)
- Renal tubular acidosis
- Some amino acidaemias

transferase deficiency and 3-oxothiolase deficiency) should be considered, as these can be easily missed. Urine organic acid profiling followed by enzyme and mutation analysis will be needed to confirm a diagnosis of these rare conditions.

It will be apparent from Box 24.2 that the predominant causes of metabolic acidosis in IMDs are ketosis and lactic acidosis. Lactic acidosis is frequently present in the acutely ill, when circulatory collapse results in tissue hypoxia. It is also frequently associated with organic acid disorders, such as propionic acidaemia and methylmalonic aciduria, as a result of secondary interference with the metabolism of coenzyme A. Ketone analysis, again, provides clues to the differential diagnosis, since, although ketosis is a consistent finding in many IMDs, it is generally absent in lactic acidosis secondary to hypoxia. Secondary causes of lactate accumulation should be excluded before an inherited disease of lactate/pyruvate metabolism is sought.

Respiratory alkalosis is a far less frequent finding in IMDs and when it is present, plasma ammonia should be measured urgently. Ammonia stimulates the respiratory centre, causing hyperventilation, resulting in a decrease in $p\text{CO}_2$. The finding of acute encephalopathy with respiratory alkalosis should raise the suspicion of hyperammonaemia that, if confirmed, should prompt immediate investigations to exclude urea cycle disorders. In hyperammonaemia due to organic acid disorders, the acid–base disturbance is typically a metabolic acidosis.

Blood glucose

Hypoglycaemia is a presenting feature of many IMDs, such as those in which there is a primary block in glucose metabolism, for example glycogen storage diseases, or those in which glucose metabolism is affected secondarily, for example tyrosinaemia type I. Further investigations to identify the underlying cause should ideally be carried out when the child is hypoglycaemic. If this window of opportunity is missed, it is often difficult to make the diagnosis, since biochemical abnormalities may not be present during normoglycaemia; the child may then have to be subjected to a controlled fast with all the risks associated with this procedure. It is best practice for any hospital with a paediatric workload to

have an agreed protocol for the investigation of hypoglycaemia. This protocol should be instituted if a child has a laboratory blood glucose of <2.6 mmol/L (or point-of-care testing glucose <3 mmol/L). Best practice guidelines are available from the MetBioNet website: www.metbio.net

Plasma ammonia

It is essential to measure plasma ammonia in all children with acute or chronic encephalopathy, recurrent vomiting or hyperventilation. Suspicion of an IMD should be raised in an infant who is well at birth, with abnormal clinical features developing only after the first 24 h once feeds have been established. Clues may be gleaned from the family history: for example, where a background of early male deaths or females with episodic illness would suggest X-linked OTC deficiency.

Hyperammonaemia is often missed, particularly in the neonatal period, when the clinical presentation is non-specific and can mimic sepsis. All centres with neonatal and paediatric units should therefore be able to provide rapid analysis of plasma ammonia. Care should be taken in the sample handling, with ammonia-free tubes used for collection of blood and the separation of plasma from cells within 15 min.

Overwhelming hyperammonaemia in the neonatal period is usually due either to a urea cycle disorder (approximately 70%) or to an organic acid disorder (approximately 30%). The hyperammonaemia in organic acid disorders is a result of a deficiency in acetyl-CoA required for the synthesis of N-acetylglutamate, a preliminary step in urea synthesis. The finding of a metabolic acidosis in a patient with hyperammonaemia is suggestive of an organic acid disorder. Since the severity of hyperammonaemia does not allow a distinction between the causes, and treatment will differ depending on the underlying defect, rapid biochemical follow-up with second-line investigations is required. Analysis of urinary organic acids and orotic acid, in conjunction with plasma amino acids and acylcarnitines, is essential; results can usually be available the same day after discussion with the regional metabolic laboratory. [Figure 24.5](#) provides a flowchart for the investigation and diagnosis of neonatal hyperammonaemia.

In older children and adults, plasma ammonia should be measured in those with unexplained encephalopathy (progressive or chronic relapsing), vomiting or drowsiness, Reye-like syndrome (hyperammonaemia, hypoglycaemia and raised amino-transferases), neurological dysfunction or ataxia. Clinical illness is often episodic in the late onset group, presentation being associated with catabolism such as occurs during infections. If a complete dietary history is taken clues, such as self-selection of a low protein diet, may be revealed. In those patients with a Reye-like presentation, the possibility of a fatty acid oxidation disorder should be considered.

Liver function tests

Jaundice, accompanied by other features of liver dysfunction, is one of the most frequent presenting features of IMDs. This is not surprising, given the central role of

the liver in metabolism. Liver function tests are therefore essential first-line investigations in suspected IMDs. They should comprise measurement of bilirubin (total and conjugated), aminotransferases, alkaline phosphatase and the prothrombin time (or international normalized ratio).

Many IMDs present with liver failure, often in the neonatal period. The liver function test findings are characterized by a severe conjugated hyperbilirubinaemia and raised serum aminotransferases. The most striking of this group of disorders is classic galactosaemia (galactose 1-phosphate uridyl transferase deficiency), which typically presents after introduction of milk feeding. The principal IMDs causing neonatal liver disease are listed in [Box 24.3](#). This box does not include the inherited hyperbilirubinaemias, which are not associated with other features of hepatic dysfunction.

Measurement of ketones

Ketosis is a normal response to fasting and, when not associated with either acidosis or hypoglycaemia, should, in infancy and childhood, be considered as physiological. However, as previously mentioned, the presence of ketones in a neonate is abnormal and requires urgent follow-up. Although ketosis is a physiological response, it is likely to be of clinical significance when it is associated with acidosis. The detection of urine ketones, using a simple point-of-care testing device, is therefore the starting point for the investigation of metabolic acidosis (see [Box 24.2](#)).

The absence of ketones can also give a clue to the underlying IMD, the most classic example of which is hypoketonaemia with hypoglycaemia due to a fatty acid oxidation defect such as MCADD. Patients with this group of disorders are able to mobilize fat stores during periods of fasting or catabolic stress but, owing to enzyme deficiencies, are unable to oxidize the fatty acids completely, leading to a relative deficiency of acyl-CoA required for ketogenesis. Although classically considered to lead to a complete absence of ketone production, it is more common to see some ketones present but at an inappropriately low concentration. The presence of ketonuria in hypoglycaemia should therefore not preclude investigation for fatty acid oxidation defects.

Urinary reducing substances

The detection of reducing substances in the urine, using simple but non-specific tablets, has historically been a mainstay of first-line testing for IMD. The withdrawal from the market of Clinistest® (Bayer) in 2011 has prompted a review of this practice and a streamlining of investigations of disorders of monosaccharide metabolism. Thin layer chromatography (TLC) may be useful to identify sugars that are present, but may give a negative result in children with IMD in whom the abnormal pathway is not stressed, for example if an infant with classic galactosaemia has been placed on a lactose-free milk formula. There are also analytical concerns with using

Key investigations: confirm hyperammonaemia with free-flowing sample (arterial if necessary)
 venous blood gases
 amino acids (plasma)
 organic acids (urine)
 acylcarnitines
 liver function tests including clotting (to exclude liver failure)

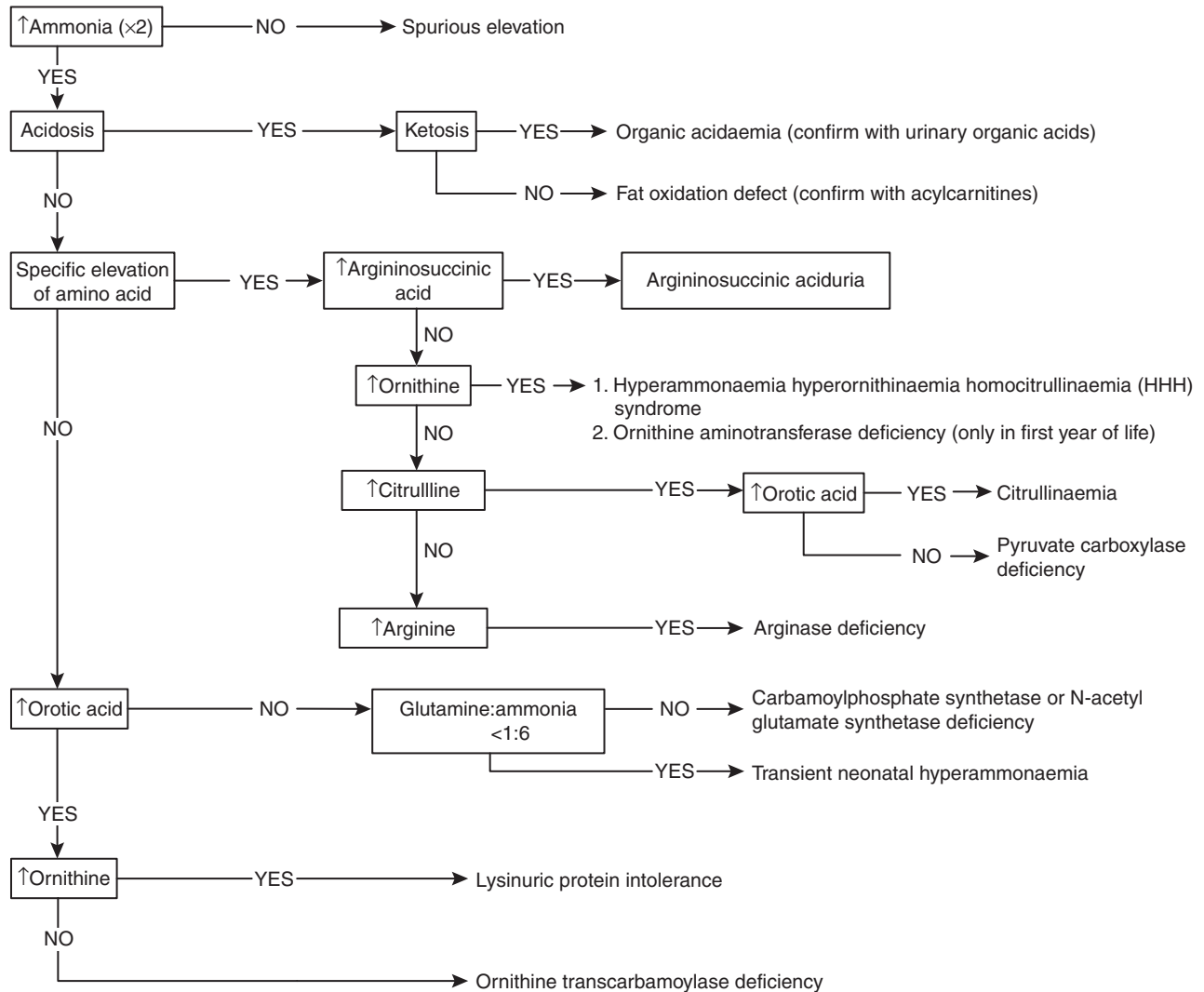


FIGURE 24.5 ■ A flowchart for the investigation of hyperammonaemia.

BOX 24.3 Inherited metabolic diseases presenting with neonatal liver disease

- Galactosaemia
- Tyrosinaemia type I
- α_1 -Antitrypsin deficiency
- Neonatal haemochromatosis
- Fatty acid oxidation disorders
- Respiratory chain disorders
- Disorders of bile acid metabolism
- Peroxisomal disorders

sugar TLC, most notably a lack of a robust external quality assurance scheme and the reproducibility of results being highly operator dependent.

In the investigation of classic galactosaemia, where jaundice and liver dysfunction can rapidly deteriorate to hepatic failure, whole blood should be sent for urgent analysis of galactose 1-phosphate uridyl transferase activity. In older infants, with progressive liver dysfunction and hepatomegaly presenting after weaning, the diagnosis of hereditary fructose intolerance should be considered. Liver biopsy for aldolase activity and mutation analysis will be required for diagnosis.

Second-line investigations

Plasma and urinary amino acids

Amino acids are important intermediates in a number of metabolic pathways and are a significant source of energy, particularly during fasting. Their metabolism may involve potentially toxic intermediates that are normally further metabolized, but may accumulate in metabolic diseases involving amino acid metabolism and thus lead to organ damage. Amino acid analysis is therefore a key second-line investigation when an inherited metabolic disease is suspected (Box 24.4), and is frequently requested as part of a metabolic screen.

Individually, amino acid disorders are rare, with the most common (PKU) having an incidence of 1 in 10 000. However, the combined incidence of all disorders that can be diagnosed using amino acid analysis is in the order of 1 in 6000. As a group, they are heterogeneous, both in terms of clinical symptoms and age of presentation. The onset of symptoms may occur anywhere from the neonatal period to adulthood, but is classically associated with periods of protein catabolism, such as delayed feeding in a neonate, or intercurrent infections in an infant. Acute symptoms are associated with the accumulation of large amounts of amino acids that overwhelm a defective pathway, resulting in the production of toxic metabolites. There are certain presenting features in neonates, e.g. acute encephalopathy and ketosis, which should prompt consideration of amino acid analysis. However, the clinical presentation of many of these disorders is disease specific, e.g. gyrate atrophy in ornithine aminotransferase deficiency, and may warrant specific investigation. An understanding of the nuances of presentation in amino acid disorders is the key to making a diagnosis.

Amino acid disorders can be simply categorized as primary or renal disorders (Table 24.2). In primary disorders, there is a defective metabolic pathway with accumulation of specific amino acids or metabolites as a result of an enzyme deficiency. A classic example is maple syrup urine disease, caused by a deficiency in the enzyme branched chain oxoacid dehydrogenase, which leads to the accumulation of valine, leucine and isoleucine and the presence of the abnormal metabolite allo-isoleucine. In primary amino acid disorders, the most informative sample is plasma, giving a snap shot of the metabolic pathway. Urine is less useful in this group since the excretion of amino acids is much more variable

BOX 24.4 Indications for plasma or urine amino acid analysis

- Neonatal lethargy, coma, seizures or vomiting
- Hyperammonaemia
- Hypoglycaemia
- Ketosis
- Metabolic acidosis or lactic acidemia
- Metabolic decompensation or encephalopathy
- Unexplained liver disease
- Unexplained developmental delay
- Renal disorders, e.g. calculi, tubulopathy

TABLE 24.2 Renal amino acid disorders

Disorder	Urinary amino acid pattern
Cystinuria	Cystine, ornithine, arginine, lysine
Hartnup disease	Neutral amino aciduria
Lysinuric protein intolerance	Lysine, ornithine, arginine
Iminoglycinuria	Proline, hydroxyproline, glycine

(especially in prematurity) and is significantly prone to interference from medication. Urine is, however, essential for the diagnosis of renal amino acid disorders such as cystinuria. In this group, amino acid metabolism is intact, but renal tubular reabsorption of specific amino acids is defective. The abnormal pattern is therefore only observed in urine.

It is important to be aware of the type of investigation undertaken when amino acid analysis is requested. Simple urine 'spot tests', such as the cyanide-nitroprusside test for cystine and homocystine, are often employed by local hospitals as a preliminary to referral to a specialist centre. These simple assays are becoming increasingly obsolete and are limited by a lack of sensitivity and specificity and should not be used in isolation. Other screening techniques may be employed, such as thin layer chromatography, but again are limited in the detection of anything but gross abnormalities: some disorders, particularly those with mild biochemical changes, may be missed by these methods. If these qualitative techniques are employed, their limitations must be understood, and conveyed to the requesting clinician, so that diagnoses are not missed.

Amino acids are best analysed using quantitative techniques such as ion exchange chromatography. These methods are sufficiently sensitive to detect the newly recognized amino acid disorders characterized by low concentrations of amino acids, such as 3-phosphoglycerate dehydrogenase deficiency, a disorder of serine synthesis.

Ion exchange chromatography is most suited to 'routine' analysis, but is limited by the long analysis time in the acute situation. To complement the approach, rapid tandem mass spectrometry (TMS), with analysis of a targeted list of amino acids and other metabolites (see section on acylcarnitines below), allows a diagnosis to be made swiftly and treatment to be instituted in an acutely sick child. Combining rapid targeted TMS testing in the acute situation, with a routine approach that clearly states which disorders are excluded, may provide a balanced solution to the provision of amino acid analysis.

Urinary organic acids

Urine organic acid profiling is an important second-line investigation in which up to 150 different IMDs can be identified from a single analysis. Organic acid analysis can identify intermediary metabolites from most amino acid, carbohydrate, purine and pyrimidine, neurotransmitter, cholesterol and fatty acid pathways,

making it a powerful investigative tool. Using organic acid analysis, metabolic disorders can be identified by the presence of pathological concentrations of normal metabolites, for example fumaric acid in fumarase deficiency, or the presence of pathological (and in some cases pathognomic) metabolites such as succinylacetone in tyrosinaemia type I. Most laboratories use gas chromatography-mass spectrometry (GCMS) to undertake organic acid analysis, providing a qualitative interpretation of the profile. However, the use of quantitative analysis by stable isotope dilution should be considered when small concentrations of critical metabolites may lead to a diagnosis.

Many disorders can be detected by organic acid analysis regardless of whether the sample is collected when the child is acutely ill or not. However, some IMDs are clearly identifiable only during episodes of acute decompensation. Medium chain acyl-CoA dehydrogenase deficiency is one such disorder where the classic pattern of a medium chain dicarboxylic aciduria (adipic, suberic and sebacic acids), with an inappropriately low excretion of ketones in the presence of the pathognomic metabolite hexanoylglycine, may only be apparent when the child is acutely unwell. Samples collected during episodes of acute decompensation are therefore vital for investigation of IMDs. The indications for urinary organic acid analysis are summarized in [Box 24.5](#).

Urinary orotic acid

Increased urinary excretion of orotic acid is a feature of four IMDs involving the urea cycle (ornithine transcarbamoylase deficiency, citrullinaemia, argininosuccinic aciduria and argininaemia) (see [Fig. 24.1](#)) and is helpful in their diagnosis. There are a number of analytical techniques that can be used for identification of this key metabolite, for example GCMS or TMS, and it is often required as part of the rapid investigation of a sick infant.

Blood acylcarnitines

With the advent of TMS technology, acylcarnitine profiling has become a crucial investigation in the field of IMDs. The use of TMS for acylcarnitine profiling, combined with simultaneous analysis of other intermediary metabolites, enables the diagnosis of many of the treatable amino acid, organic acid, fatty acid oxidation

and urea cycle disorders. In a clinical emergency, if one of these disorders is considered, then discussion with a specialist metabolic service to arrange urgent analysis should be a priority.

Acylcarnitine profiling allows the quantitation of saturated and unsaturated acylcarnitine species required for the diagnosis of a number of IMDs. In addition, the determination of both free and total carnitine (i.e. overall carnitine status) is required for the investigation of suspected carnitine transporter defects. Typically, acylcarnitine analysis is undertaken on plasma or blood spot samples, with only a small sample usually being necessary. Bile samples, collected post-mortem, on to newborn screening cards, can be used if the suspicion of IMD is raised at autopsy.

It should be noted that some conditions can only be diagnosed confidently with the use of acylcarnitine profiling: an example is VLCADD. Acylcarnitine analysis should therefore be undertaken if a fatty acid oxidation defect is suspected; these disorders typically present with hypoketotic hypoglycaemia, but may be associated with cardiomyopathy, rhabdomyolysis or hepatomegaly. In a pregnant woman with AFLP or HELLP syndrome, the possibility of LCHADD in her infant should be considered. Rapid acylcarnitine profiling is required for the newborn infant of a mother with a history of either condition.

Blood lactate and pyruvate

Lactate and pyruvate are important intermediaries in a number of metabolic pathways, particularly in the production of ATP under anaerobic conditions. It is therefore not surprising that assessment of these pathways, by analysis of blood lactate, is an essential measurement in the investigation of a suspected IMD. In the absence of tissue hypoxia, and assuming the sample was collected properly, using an appropriate antiglycolytic agent, a blood lactate of >3 mmol/L should be considered abnormal and further investigations for an IMD initiated.

Lactic acidemia is found in many groups of IMDs (see [Box 24.1](#)) and may lead to a metabolic acidosis when the lactate concentration exceeds 5 mmol/L. It is a frequent finding in organic acidemias, urea cycle disorders and fatty acid oxidation disorders, as a result of secondary interference with the metabolism of coenzyme A, and is often accompanied by ketosis. These groups of IMDs can be rapidly distinguished with follow-up investigations such as organic acid, amino acid and acylcarnitine analysis.

The primary lactic acidemias, for example defects in pyruvate metabolism or respiratory chain function, often present with severe metabolic acidosis in the neonatal period. A distinguishing feature of this group, as opposed to the secondary IMD groups noted above, is that the clinical picture is unrelated to intake of protein. In the primary lactic acidemias, analysis of both blood lactate and pyruvate, calculation of the lactate:pyruvate (L:P) ratio and measurement of cerebrospinal fluid (CSF) lactate are of assistance in directing further investigations. It must be noted that pyruvate is

BOX 24.5 Indications for urinary organic acid analysis

- Unexplained metabolic acidosis
- Hypoglycaemia
- Ketonuria (urgent in newborn)
- Lactic acidosis
- Hyperammonaemia
- Encephalopathy
- Unexplained neurological disorder
- Multisystem failure
- Hepatic dysfunction

an extremely unstable analyte and needs careful sample handling with immediate precipitation of whole blood. However, if accurate measurement of lactate and pyruvate is undertaken, the L:P ratio can give valuable insight into the cytoplasmic redox status. An elevated L:P ratio (>25) is suggestive of a respiratory chain defect or pyruvate carboxylase deficiency, with a normal ratio (<25) suggesting a disorder of gluconeogenesis or pyruvate dehydrogenase deficiency.

Further information can be gained from assessing plasma lactate concentrations in relation to the fed and fasted states. This is of particular use in the investigation of glycogen storage disorders (GSDs). Lactic acidemia in the fasted state is a key finding in GSD type I (glucose 6-phosphatase deficiency), in contrast to GSD type III (amylo-1,6-glucosidase deficiency), in which the increase in lactate is most noticeable after feeding. In investigating these conditions, it is important that samples are collected both pre- and postprandially so that lactic acidemia is not missed. It is for this reason that lactate is a key measurement in many diagnostic fasts and 24h metabolite profiles (see below).

Urinary glycosaminoglycans

The mucopolysaccharidoses are a group of lysosomal storage disorders characterized biochemically by the accumulation of glycosaminoglycans (GAGs) in the urine. Clinically, they present with coarsening of the facies, in conjunction with features including skeletal abnormalities, hepatosplenomegaly and hearing loss. The majority of these features are not present at birth but develop progressively with age. If this group of disorders is suspected, a random urine sample should be collected for analysis of GAGs. In the first instance, quantitative analysis is usually undertaken with interpretation against age-related reference ranges. This is then followed-up by electrophoresis to identify the individual GAGs from the typical patterns characteristic of each disorder (see Table 24.3). Definitive diagnosis is by enzyme analysis.

Plasma very long chain fatty acids

Very long chain fatty acids, with carbon chain length ≥ 22 , are exclusively metabolized by β -oxidation in peroxisomes. The analysis of plasma VLCFAs is therefore an

important second-line investigation in suspected peroxisomal disorders. These are a genetically heterogeneous group but, in disorders both of peroxisomal biogenesis (such as Zellweger syndrome), and single peroxisomal protein defects (the most common of which is X-linked ALD), β -oxidation is impaired, leading to the accumulation of saturated VLCFAs. The finding of an elevated ratio of C26:C22 fatty acids is a clear indication of peroxisomal dysfunction. However, it should be noted that although analysis of VLCFAs identifies the presence of a peroxisomal disorder, results should be viewed in conjunction with the clinical findings, imaging and further laboratory analysis, such as erythrocyte plasmalogen, bile acid intermediates and pristanic acid to enable a precise diagnosis.

Functional and loading tests

The aim of functional or loading tests is to unmask IMDs that may be difficult to diagnose by conventional techniques. They are most often used in situations where metabolic findings are not clearly abnormal and the diagnosis is uncertain, for example many of the fatty acid oxidation disorders, where biochemical findings may be subtle, or absent, when the patient is well. If the window of opportunity to collect samples when the patient is hypoglycaemic or acutely unwell is missed, then a controlled diagnostic fast may be carried out to aid the diagnosis.

Diagnostic fast

During a diagnostic fast, blood samples are collected hourly. Investigations should include markers of intermediary metabolism (non-esterified fatty acids, hydroxybutyrate and lactate), acylcarnitines and counter-regulatory hormones (insulin, growth hormone and cortisol). Urine organic acids and blood acylcarnitine should be measured at the beginning and end of the fast. The blood glucose concentration is monitored using a point-of-care testing device appropriate for the detection of hypoglycaemia; such results must be confirmed in the laboratory. Given the risks associated with this procedure, diagnostic fasts must be performed under medical supervision in a specialist unit. In addition to providing valuable diagnostic information, a fast allows the determination of the

TABLE 24.3 Mucopolysaccharidoses

Type	Disorder	Urinary glycosaminoglycan	Enzyme deficiency
I	Hurler	Dermatan sulphate, heparan sulphate	α -Iduronidase
II	Hunter	Dermatan sulphate, heparan sulphate	Iduronate sulphatase
IIIA	Sanfilippo A	Heparan sulphate	Heparin sulphamidase
IIIB	Sanfilippo B	Heparan sulphate	N-acetyl- α -glucosaminidase
IIIC	Sanfilippo C	Heparan sulphate	N-acetyltransferase
IIID	Sanfilippo D	Heparan sulphate	N-acetylglucosamine sulphatase
IVA	Morquio A	Keratan sulphate	N-acetylgalactosamine 6-sulphatase
IVB	Morquio B	Keratan sulphate	β -Galactosidase
VI	Maroteaux-Lamy	Dermatan sulphate	N-acetylgalactosamine 4-sulphatase
VII	Sly	Dermatan sulphate, heparan sulphate, chondroitin sulphate	β -Glucuronidase

maximum safe fasting period before the onset of hypoglycaemia, and is often undertaken on a yearly basis in patients with glycogen storage disease.

Allopurinol loading test

Loading tests can be used to expose an enzyme defect, where there is residual activity. The allopurinol test can be used in this way to confirm heterozygote status in female carriers of OTC deficiency. Urine is collected for measurement of orotidine and orotic acid, before and after administration of a dose of allopurinol. Although this is a safe and noninvasive approach, not all at-risk females will be detected. This strategy, although still in use, is now being superseded by the use of mutation analysis.

Confirmatory investigations

Enzyme analysis: general principles

Final confirmation of the diagnosis of a suspected IMD often requires enzyme analysis. In some cases, confirmatory testing is not undertaken, because the diagnosis can be inferred from a typical clinical history and a characteristic pattern of abnormal metabolites. In conditions such as PKU, the use of an invasive procedure, such as liver biopsy, to obtain tissue for enzyme analysis would not contribute further to the management of the patient. However, the majority of patients with a suspected IMD will have enzyme analysis to confirm the diagnosis.

Enzyme analysis can be undertaken in various tissue types, ranging from blood cells to liver and muscle samples, obtained by biopsy. It is of particular importance that for IMDs that are tissue specific, an appropriate specimen is used, e.g. liver biopsy samples in the urea cycle disorders and muscle or liver in the glycogen storage disorders that are either muscle or liver specific. If a patient with a suspected IMD is not expected to survive, plasma, urine, skin and, if possible, muscle and liver biopsies should be collected. Skin biopsies are of particular use since fibroblast cultures can be established and DNA harvested if required.

Red cell galactose 1-phosphate uridylyltransferase. Classic galactosaemia, due to a deficiency in the enzyme galactose 1-phosphate uridylyltransferase (Gal-1-PUT or GALT), leads to the accumulation of galactose and other toxic metabolites on introduction of lactose into the diet. In the past, testing for urine reducing substances provided an easily available screening test for galactosaemia. Now, clinical suspicion should be urgently followed-up with confirmatory testing of red cell GALT activity. The requesting clinician must state clearly the transfusion status of the child, since previous blood transfusion will invalidate the result. If a child has been transfused, it may be possible to infer the diagnosis by analysing parental blood and demonstrating that they are heterozygotes.

Lysosomal enzyme screening. The lysosomal storage disorders are a group of approximately 40 different

BOX 24.6 Clinical and laboratory features suggestive of a possible lysosomal storage disorder

- Neonatal ascites
- Hydrops fetalis
- Coarse features
- Skeletal dysplasia
- Hepatosplenomegaly
- Neurodegeneration
- Corneal clouding
- Cherry red spots in macula
- Vacuolated lymphocytes

IMDs characterized by the accumulation of macromolecules in lysosomes, as a result of specific enzyme or protein transporter defects. The accumulated material results in an increase in the size and number of these organelles, leading to cellular dysfunction and subsequent pathological features. The lysosomal storage disorders are generally classified on the basis of the accumulated material, for example sphingolipidoses, mucopolysaccharidoses and mucopolysaccharidoses. These conditions typically present in a chronic and progressive manner, in many cases with neurological degeneration. The principal clinical features are summarized in [Box 24.6](#).

Diagnosis of this group of disorders is often difficult, as there is an overlap in the clinical presentation of each disease. A 'screening' approach is, therefore, adopted by specialist laboratories: this ensures the maximum possibility of making a diagnosis. Groups of lysosomal enzymes in plasma and leukocytes are analysed depending on the clinical features, for example whether the patient has neurodegeneration or the presence or absence of hepatosplenomegaly. The choice of enzymes is tailored to exclude the majority of lysosomal storage disorders. This approach requires good communication between the local hospital and specialist enzyme laboratory so that a systematic approach to the investigation is undertaken.

Complementation studies

Although in the vast majority of IMDs the enzyme deficiency is well characterized, there are still a number of disorders in which the protein or gene defect has not been identified. In these conditions, such as those that lead to decreased synthesis of adenosylcobalamin and the clinical features of cobalamin-responsive methylmalonic aciduria (classified as Cbl_a and Cbl_b), a different approach to confirm the diagnosis may be taken. To distinguish these different groups, complementation studies can be undertaken in cultured fibroblasts. The principle of the investigation is to compare the incorporation of a marker substance, in this case propionate, in parallel fused and unfused cell cultures. The classification of the patient in this way is of use in assessing the possible prognosis and excluding disease in other family members. However, it is likely that this approach will be superseded once the enzyme or gene locus is identified.

Genetic mutation analysis

Mutation analysis is playing an increasingly important role in the diagnosis of IMDs. The use of next generation sequencing has enabled a rapid and more cost-effective approach and is becoming a mainstay of diagnostic pathways.

Mutation analysis has many benefits in the diagnosis of IMDs. It is of particular use in prenatal diagnosis, where mutation analysis in chorionic villus samples (CVS) can be performed much earlier than enzyme diagnosis in cultured amniocytes. Once the mutation has been defined in the proband, testing of family members may overcome the need for invasive sampling, such as liver biopsies or functional tests that may give ambiguous results.

Where mutations are classified and the genotype/phenotype correlation is well recognized, this may provide useful information regarding prognosis and possible treatment options. In some disorders, such as MCADD, a common disease-causing mutation (A985G) is present in many cases, allowing the use of mutation analysis to secure the diagnosis. The approach of combining metabolite assays with mutation analysis confirms the diagnosis rapidly and may obviate the need for fibroblast culture. Problems may arise, however, when compound heterozygotes are identified: in these instances it might not be clear what the clinical outcome of a combination of potentially less severe mutations would be.

PRENATAL DIAGNOSIS

The availability of prenatal diagnosis gives a number of options to a family in whom a serious genetic disorder is suspected. In the field of IMDs, prenatal diagnosis is often requested if a previously affected child has a poor prognosis, particularly where the options for treatment are limited. Inherited metabolic diseases are generally diagnosed prenatally using either enzyme analysis in cultured amniocytes or mutation analysis in CVS.

Amniocentesis, undertaken in either the first or second trimester, provides fetal amniocytes for culture. Enzyme analysis can be undertaken in these cultures and the activity compared with appropriate reference ranges

to confirm the diagnosis. The limitation of this approach is the time scale: amniocyte culture typically takes 4–6 weeks to establish, leading to additional anxiety for the family and limited time for decision-making. Since only a small number of analyses are undertaken each year, prenatal diagnosis of IMDs using these techniques is undertaken at only a few specialized centres.

Chorionic villus sampling is generally undertaken at 9–12 weeks of gestation, providing the opportunity for earlier diagnosis and giving additional time for counselling and safer termination, if this option is chosen. In this technique, fetal villus cells obtained by amnioscopy are dissected from the maternal decidual tissue, and DNA extracted and used for mutation analysis. Adherence to strict techniques is paramount, to avoid maternal cell contamination that may result in diagnostic ambiguity.

MANAGEMENT

Many treatment modalities are adopted to manage IMDs, dependent on whether pathology results from deficiency of the product, the build up of a toxic metabolite or a combination of the two. At present, most treatments remain symptomatic rather than curative.

Strategies to replace a missing product

Supply of precursor

Enzymes catalyse reactions that would proceed at a much reduced rate in their absence. Provision of large doses of precursor can shift the reaction in the direction of making the product by a mass action effect: in some cases this is sufficient to ameliorate symptoms.

Hydroxycobalamin is converted in the body via a number of intracellular conversions from its dietary hydroxycobalamin form to its two active forms, adenosylcobalamin and methylcobalamin. In the cobalamin disorders, blocks in these conversions lead to accumulation of homocysteine, methylmalonate or both, dependent on the position of the block (Fig. 24.6). Cobalamin C disorder is the commonest of these rare IMDs. Patients present in the neonatal period and first year with neurological deterioration, poor feeding and hypotonia, and often develop

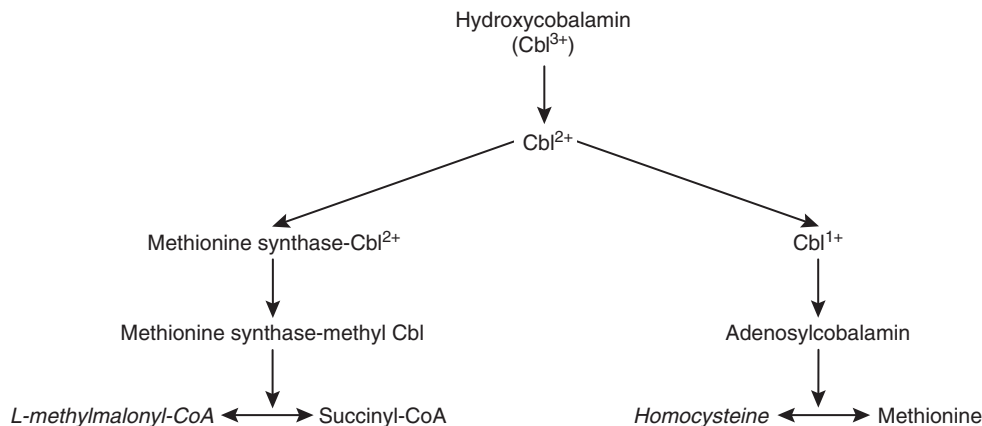


FIGURE 24.6 ■ Cobalamin metabolism.

seizures. Anaemia is common, as is multiorgan involvement including the kidneys, liver and heart. The diagnosis is suggested by elevated plasma homocysteine and urinary methylmalonate concentrations, in the absence of hydroxycobalamin deficiency. Diagnosis is confirmed by fibroblast incorporation studies and decreased synthesis of adenosyl- and methylcobalamin. Treatment with regular intramuscular hydroxycobalamin leads to a marked reduction in homocysteine and methylmalonate, but not complete biochemical normalization. Clinical improvement is limited even in prospectively treated affected siblings.

Replacement of product

Replacement of the deficient product of a blocked pathway is another solution to the treatment of some IMDs: it effectively bypasses the block. Its efficacy depends on the product being available and its potential to be given in a form that can be taken by the patient and then reach its site of action in a usable form.

Such an approach is used in the congenital disorder of glycosylation (CDG) Ib (phosphomannose isomerase deficiency). Glycosylation is an essential process in the body, as almost all proteins that are secreted or membrane bound are dependent on their glycans (carbohydrate side chains) for normal activity. The production of glycoproteins begins with the manufacture of the oligosaccharide precursor and its transfer to the nascent polypeptide chain in the endoplasmic reticulum, with subsequent modification within the Golgi apparatus to produce the desired glycan. Defects of the pathway were first established when examining transferrin, which has two glycans with four sialic acid residues. Defective glycosylation results in variation in the number of sialic acid residues, which affects the charge on the protein. This can be detected using isoelectric focusing, which shows additional bands. Children with CDG Ib present with recurrent vomiting and diarrhoea, coagulopathy, protein-losing enteropathy, hypoalbuminaemia, delayed growth and hyperinsulinaemic hypoglycaemia. Liver fibrosis has also been noted. Treatment with oral mannose effectively bypasses the block by replacing the missing product (Fig. 24.7). Clinically, there is a reduction in diarrhoea and vomiting with improved growth. Biochemically, the isoelectric focusing pattern normalizes

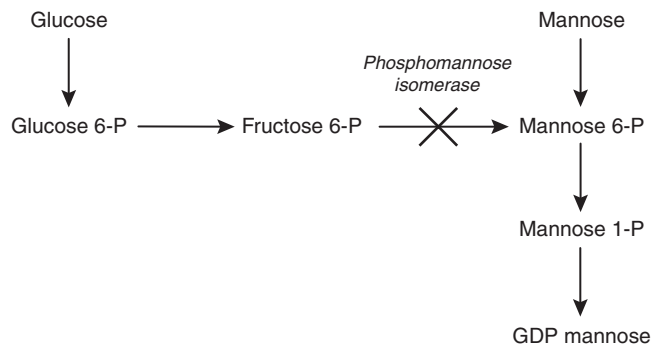


FIGURE 24.7 ■ Mannose therapy for CDG Ib (phosphomannose isomerase deficiency). Mannose is phosphorylated by hexokinase to mannose 6-phosphate, thereby bypassing the block. GDP, guanosine 5'-diphosphate; P, phosphate.

as glycan production beyond the block is normal. Plasma concentrations of antithrombin III, the cause of the coagulopathy, also return to normal.

In dihydropteridine reductase deficiency, there is decreased recycling of biopterin, the essential cofactor for phenylalanine hydroxylase, the enzyme responsible for the conversion of phenylalanine into tyrosine, with resultant phenylketonuria. In addition, biopterin is the essential cofactor for two other hydroxylases in the body, namely tyrosine and tryptophan hydroxylases, required for the production of L-DOPA and 5-hydroxytryptophan (5-HT), respectively. These are, in turn, converted to dopamine and serotonin, which are two of the major neurotransmitters in the brain. Dihydropteridine reductase deficiency was originally termed malignant PKU, as following the introduction of the PKU screening in the 1970s it became apparent that there were very rare patients who still went on to have marked developmental delay and characteristic axial hypotonia and limb dystonia, in spite of excellent phenylalanine control by diet. Now, all neonates with hyperphenylalaninaemia detected on newborn screening are screened for biopterin defects, usually performed on dried blood spots by measuring total biopterins. Dihydropteridine reductase deficiency is managed with a phenylalanine restricted diet and supplementation with 5-HT and L-DOPA. The latter is given in combination with carbidopa in the ratio 4:1. Carbidopa is a peripheral decarboxylase inhibitor, which reduces peripheral conversion of L-DOPA to dopamine, thereby increasing the amount of L-DOPA reaching the brain and reducing side-effects, which include nausea, vomiting, postural hypertension, dystonic reactions and psychological disturbance. Dose increments should be small, as initially patients are very sensitive to the effect of the treatment owing to upregulation of receptors. Dosage requirements are monitored clinically and by measuring the breakdown products of dopamine and serotonin, namely homovanillic acid (HVA) and 5-hydroxyindoleacetic acid (5-HIAA), in the CSF. A slightly greater dose of L-DOPA is given compared with 5-HT, to mimic the naturally slightly higher amounts within the brain. In addition, folinic acid is required to combat the associated central folate deficiency seen in this condition.

Replacement of the missing product is also used in pyridoxal phosphate-dependent seizures in pyridoxamine 5-phosphate oxidase deficiency. Pyridoxal phosphate is the active form of pyridoxine, required for a number of essential reactions in the body and particularly in the brain. Pyridoxamine 5-phosphate oxidase deficiency presents with devastating neonatal seizures, which can be effectively treated by giving pyridoxal phosphate orally. The cessation of seizures is immediate and dramatic but associated with transient hypotonia and apnoea. Treatment should therefore be initiated with full resuscitation facilities available.

Replacement of the missing product in utero has been successful in 3-phosphoglycerate dehydrogenase deficiency, a serine biosynthesis disorder, characterized by congenital microcephaly, severe psychomotor retardation and intractable seizures. The treatment of affected children with oral L-serine has proved effective in preventing the seizures: however, development is not normalized and

the congenital microcephaly suggests impaired fetal brain growth in utero. It has now been shown that giving L-serine to the mother during pregnancy can reverse the decrease in fetal head growth in utero with resultant normal neurological development. The key diagnostic investigation for this condition is analysis of serine concentration in CSF.

Synthetic analogues

It may not be possible to replace the product directly, owing to difficulties with absorption and denaturation when exposed to the acidity of the stomach and digestive enzymes. Synthetic analogues of the missing product are, however, used in some circumstances. N-acetylglutamate synthetase (NAGS) deficiency is a rare urea cycle defect that may present with devastating hyperammonaemic crises. N-acetylglutamate is an essential activator for carbamoylphosphate synthetase, the second enzyme in the urea cycle pathway (see Fig. 24.1). Carbamylglutamate, a synthetic analogue of N-acetylglutamate, is an effective treatment for NAGS deficiency. It allows activation of carbamoylphosphate synthetase, in the absence of a working NAGS enzyme, reducing the risk of hyperammonaemia and improving neurological outcome. In organic acidaemias, hyperammonaemia can occur as a result of secondary inhibition of NAGS; carbamylglutamate is used in their management.

Receptor agonists can be used where the site of action of a deficient product is a receptor. Aromatic amino acid decarboxylase (AADC) converts 5-HT into serotonin and L-DOPA into dopamine. Clinical features of AADC deficiency include intermittent oculogyric crises, limb dystonia, generalized hypotonia, developmental delay and autonomic dysfunction. The diagnosis is indicated by low concentrations of the neurotransmitter breakdown products, HVA and 5-HIAA, in CSF. Dopamine receptor agonists, such as pramipexole and rotigotine, are being used to reduce the frequency of crises, and also improve voluntary movements, but the benefits have so far been variable.

Alternate product

A further strategy employed in the treatment of IMDs is to use an alternate product to bypass the block. The classic example is the use of a ketogenic diet in deficiency of glucose transporter type 1 (GLUT1), a membrane transporter that transports glucose across the blood–brain barrier: a defective protein results in hypoglycaemia in the brain, with a normal plasma glucose concentration. Presentation is in infancy with epilepsy, global developmental delay and complex movement disorders. Diagnosis requires the simultaneous measurement of glucose in the plasma and CSF which reveals a decreased CSF:plasma glucose ratio. Although glucose is the preferred fuel for the brain, ketones may be used for energy and gain entry to the brain via the monocarboxylate cotransporter (MCT1) transporter, thus avoiding the block (Fig. 24.8). The ketogenic diet promotes ketogenesis by restricting carbohydrate (typically a 4:1 ratio of fat to combined protein and carbohydrate), with clinical improvement in seizure control, development and movements. The condition tends to stabilize after puberty, but the degree of residual impairment is highly variable.

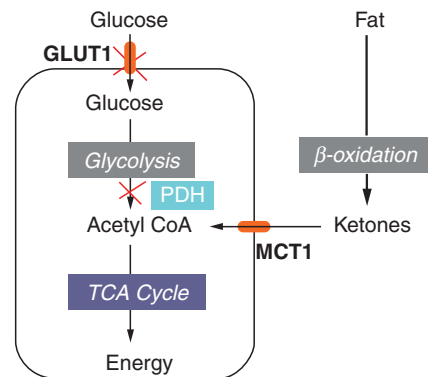


FIGURE 24.8 ■ Providing alternate fuel utilizing ketones in the management of GLUT1 deficiency and pyruvate dehydrogenase deficiency. PDH, pyruvate dehydrogenase; GLUT1, glucose transporter protein 1; MCT1, monocarboxylate transporter 1; TCA cycle, transcarboxylic acid cycle.

Ketogenic diets are similarly used in pyruvate dehydrogenase deficiency, bypassing glycolysis.

Inhibition of product breakdown

Enzyme blocks in IMDs are not usually complete, and usually some product is made. Preventing or delaying breakdown of the product can have useful therapeutic effects. In AADC deficiency described above, a combination of a dopamine agonist and a monoamine oxidase inhibitor, such as tranlycypromine, can be used in combination. The latter prevents the natural breakdown of dopamine and serotonin and therefore boosts their concentrations. This has the positive effect of reducing the frequency of oculogyric crises and improving muscular tone.

Enzyme replacement therapy

Replacement of the missing enzyme is an obvious potential solution to the management of IMDs, but many obstacles have to be overcome, including production, purification and targeting of the enzyme. Access to the CNS across the blood–brain barrier remains a technical challenge. Nearly 40 years ago, it was noted that coculture of fibroblasts from Hurler and Hunter syndrome patients mutually corrected the defects in vitro, suggesting the existence of a diffusible substance that could correct the block. Twenty years later, the first enzyme replacement therapy (ERT) became commercially available. Gaucher disease (glucocerebrosidase deficiency) has been the forerunner of the ever expanding number of conditions considered for ERT. Storage of glucocerebrosidase occurs in liver, spleen and bone, and, in two out of the three forms, brain. Type 1 is the most common, presenting with hepatosplenomegaly, anaemia, lung infiltration and bony changes: the brain is not involved. Type 1 is prevalent in the Ashkenazi Jewish population. Type 2 is rapidly progressive with severe neurological involvement and hepatosplenomegaly. Type 3 has brain involvement, but a much milder course. Inheritance of all three forms is autosomal recessive. The diagnosis is made by measuring the glucocerebrosidase activity in leukocytes.

The absence of brain involvement in type 1 made it an obvious target for ERT as the enzyme does not cross the blood–brain barrier. The enzyme used is a recombinant modified human placental glucocerebrosidase, with an exposed mannose terminus to enhance lysosomal targeting, following uptake by macrophages. Uptake of glycoproteins is mediated via carbohydrate-specific receptors, which are present on macrophages. Enzyme replacement therapy has revolutionized the management of type 1 Gaucher disease, with improvement in haematological parameters, regression of organomegaly and reduction in bone pain. The enzyme is given fortnightly as an intravenous infusion, initially in hospital, but later at home, utilizing a home-care nursing team. Some individuals produce antibodies to the enzyme protein, but therapy can usually continue.

Fabry disease (α -galactosidase deficiency) is another condition for which ERT has been developed. Previously, treatment was supportive, including dialysis and the use of analgesics. Enzyme replacement therapy is effective in reducing plasma glycosphingolipid concentrations, preserving renal function, improving cardiac conduction and reducing pain.

Replacement enzymes have been developed for other lysosomal storage disorders, including: Hurler syndrome (α -iduronidase deficiency); Hunter syndrome (iduronate-2-sulphatase deficiency); Morquio syndrome Type A (N-acetylgalactosamine-6-sulfatase deficiency); Maroteaux Lamy syndrome (N-acetylgalactosamine-4-sulfatase deficiency); Pompe disease (acid maltase deficiency), and Wolman disease (lysosomal acid lipase deficiency). At present, relatively small numbers of patients have been treated, for relatively short time periods. It is hoped that early treatment will prevent storage occurring and therefore give optimal outcomes, but, for those conditions with neurological involvement, the lack of brain penetrance remains a concern.

Cofactor supplementation

Many enzymes have vitamin cofactors that are essential for normal function. Cofactor supplementation may increase residual enzyme activity in deficient patients. Although the absolute response may be quite small, because clinical features often only develop at extremely low levels of enzyme activity (<5%), even a small increase (e.g. by 1% of normal activity) may prove to be clinically significant.

Patients may have sensitive deficiencies that are totally correctable by giving the enzyme cofactor. Some conditions can be characterized by their response to cofactor supplementation. An example is homocystinuria (cystathionine β -synthase deficiency), which is classified as being either pyridoxine responsive or pyridoxine unresponsive. In patients who respond to pyridoxine, there is a reduction in fasting plasma homocysteine concentration and characteristic hypermethioninaemia, on pyridoxine supplementation. A pyridoxine challenge should therefore be undertaken in all new patients, with measurements of plasma homocysteine and methionine concentrations before, during and after supplementation. Some patients have a partial response to pyridoxine.

Once a diagnosis has been established, cofactors are routinely given in a number of IMDs, to assess whether any clinical benefit can be gained by supplementation (Table 24.4).

TABLE 24.4 Vitamin cofactors, supplementation of which may be beneficial in inherited metabolic disorders

Cofactor	Disorder
Ascorbic acid	Tyrosinaemia type III Transient tyrosinaemia of the newborn Glutathione synthase deficiency Congenital lactic acidosis
Biotin	Biotinidase deficiency Multiple carboxylase deficiency Propionic acidaemia Pyruvate carboxylase deficiency
Biopterin	Phenylalanine hydroxylase deficiency
Cobalamin	Methylmalonic aciduria
Folate (given as folic acid)	Dihydropteridine reductase deficiency Uridine monophosphate synthase deficiency Methylene tetrahydrofolate reductase deficiency Methionine synthase deficiency Primary hyperoxaluria type 1
Pyridoxine	Homocystinuria (cystathionine β -synthase deficiency) Ornithine aminotransferase deficiency Pyridoxine-responsive seizures
Pyridoxal phosphate	Pyridox(am)ine 5'-phosphate oxidase deficiency
Riboflavin	Glutaric aciduria type I Multiple acyl-CoA dehydrogenase deficiency
Thiamin	Congenital lactic acidosis Congenital lactic acidosis Maple syrup urine disease Pyruvate dehydrogenase deficiency
Tocopherol	Glutathione synthetase deficiency
Ubiquinone	Mitochondrial disorders

Organ transplantation

Purified enzymes for replacement therapy are available for only a very small number of IMDs. A more crude form of enzyme replacement is whole organ transplantation. This most frequently involves the liver, as this is the principal site of many metabolic pathways, for example the urea cycle. It is important to examine whether there are significant extrahepatic manifestations of the disorder for which transplantation is being considered. If there is significant enzyme expression in other tissues, then it may be that liver transplantation, although improving hepatic enzyme function, will not counteract these other harmful effects. This importance of extrahepatic morbidity is demonstrated by methylmalonic acidaemia, where liver transplantation fails to protect the patient from further neurological decompensation. In contrast, the risk of metabolic stroke post liver transplantation in propionic acidaemia appears much lower, even though the enzyme deficiencies affect the same catabolic pathway (see Fig. 24.9). Correcting the defect in the liver abolishes metabolic decompensation, thereby permitting a free diet. Cerebrospinal fluid propionate concentrations do not completely normalize and therefore it has been suggested that relaxing the diet post transplant, but continuing modest protein restriction, may abolish the risk of further neurological sequelae.

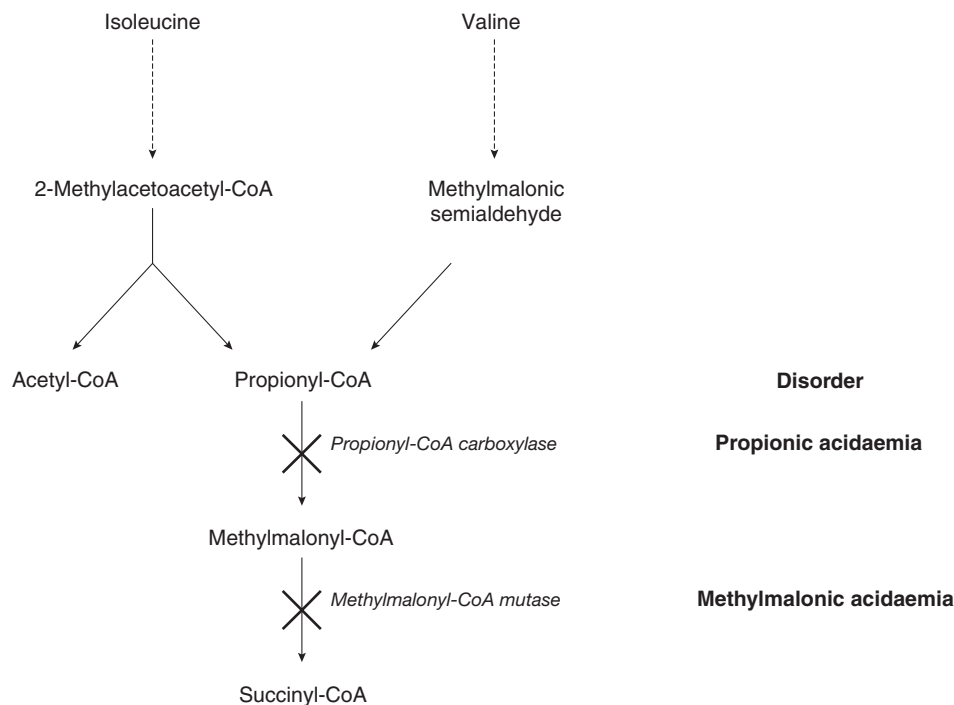


FIGURE 24.9 ■ Isoleucine and valine metabolism, showing the position of enzyme blocks in propionic acidaemia and methylmalonic acidaemia.

Liver transplantation can also be used in urea cycle defects, but the decision as to whether to transplant depends on the severity of the defect, the extent of pre-existing damage from previous hyperammonaemic crises and the future risk of longer term complications of the disease. The survival rate for liver transplantation has improved dramatically and, in children, is quoted as being >95% at five years. However, after transplantation, patients will require lifelong immunosuppressant therapy to prevent organ rejection, which increases the chance of significant sepsis and the development of lymphoproliferative disease. The use of partial liver transplantation, that is, transplanting a donor lobe without removing the native liver, has the added advantage of correcting the metabolic defect (100% liver activity is not required for normal function), while leaving the native liver (which is otherwise healthy) in situ. A partial liver transplant also has the potential, if gene therapy becomes standard practice, to allow the anti-rejection drugs to be stopped and the transplanted lobe to be rejected.

An alternative to transplantation of a whole liver, or a lobe of a liver, is hepatocyte transplantation. Neonates who present with OTC deficiency have a very poor prognosis. Mortality is extremely high and, even in prospectively treated patients (including those receiving a liver transplant in the first six months of life), neurological damage remains common. Hepatocyte transplantation has been used as a bridge to organ transplantation. Donor hepatocytes are infused into the hepatic circulation to seed in the native liver. Some of these cells engraft and replicate, thus increasing enzyme activity within the liver. This technique has been used to stabilize babies, and to avoid hyperammonaemic crises, before liver transplantation, which is usually carried out at about six months of age. Neurological outcome has been good. It is hoped that further refinement

of this technique may allow longer lasting effects that would obviate the requirement for later surgical transplantation. Conditions in which there is liver destruction are likely to benefit the most, as this provides a stimulus for hepatic regeneration to which the transfused hepatocytes respond. Hepatocyte transplantation would be appropriate for several inherited disorders including Crigler–Najjar syndrome, glycogen storage disease type Ia and infantile Refsum disease. Immunosuppression is still required, but, as with auxiliary liver transplantation, there is the potential for withdrawing immunosuppressants when gene therapy becomes feasible, as the native liver remains in situ. A further advantage of hepatocyte transplantation is that livers that might otherwise be considered unsuitable for liver transplantation can be used as a source of healthy hepatocytes, as can redundant liver lobes following the transplant of livers from adults into children.

Bone marrow transplantation has been used in a number of lysosomal disorders, fundamentally changing the natural history of these conditions. Bone marrow transplantation has an advantage over ERT, in that a small number of stem cells do pass beyond the blood–brain barrier and therefore have the ability to improve neurological outcome. Bone marrow transplantation has been attempted in nearly all lysosomal disorders, but has proven ineffective in some, particularly those in which there is severe rapid neurodegeneration, such as San Filippo syndrome (MPS III). Bone marrow transplantation has proven successful in MPS type I, which is the model lysosomal condition for bone marrow transplantation, but it needs to be performed early, ideally within the first 12–18 months of life. However, bone marrow transplantation does not ameliorate all the problems of this condition, including articular disease and spinal problems.

Gene therapy

Correction of single gene disorders by gene therapy (see also Chapter 43), replacing the faulty copy of the gene with a fully working copy, remains the ultimate goal for the treatment of IMDs. The technical challenges have proven considerable; for example the cystic fibrosis gene was first cloned at the end of the 1980s, and nearly 20 years later gene therapy remains in development.

The first condition for which gene therapy was used clinically was adenosine deaminase (ADA) deficiency, one of the causes of severe combined immunodeficiency disease (SCID). This is a combined profound impairment of both humoral and cellular immunity, presenting with multiple life-threatening recurrent infections. Bone marrow transplantation gives a good chance of complete cure; however, if no histocompatible donor is identified, ERT with polyethylene glycol-modified ADA (PEG-ADA) is given every one to two weeks, by intramuscular injection. Gene therapy trials started in the early 1990s, and involved the culture and collection of peripheral blood T cells of patients, insertion of a fully functional ADA gene, by means of a retroviral vector, and reinfusion into the patient. This improved patients' immune function, but required repeated treatments as lymphocytes have a finite life span of a few months. The technique was further developed using gene transfer into stem cells, which have an unlimited life span. However, the retroviral vector randomly inserted the gene into the cells' DNA. A total of four children subsequently developed leukaemia owing to the gene being inserted into an oncogene and therefore altering the control of this gene. Subsequent development and refinement of gene transfer technologies have allowed safer and more effective gene transfer and expression with 70% having sufficient improvement in immune function to remain off PEG-ADA. No serious adverse events have been reported with the refined vectors. Gene therapy trials are now under way for a number of IMDs.

Other molecular therapies

Nonsense mutations create a premature stop codon thus producing a truncated, usually non-functioning, protein. It is estimated that nonsense mutations account for 5–15% of disease-causing mutations. It was recognized that amino-glycosides could influence the ability to read through the premature stop codon and so produce a normal length functioning protein. Importantly, this affects only premature, and not normal, stop codons. Subsequent development of small read-through molecules has shown that boosting the affected enzyme by small amounts can significantly reduce disease severity. Clinical trials in cystic fibrosis have been promising and potential trials for patients with IMDs secondary to nonsense mutations are currently being planned.

Exon skipping is another technique to boost functioning protein production, by masking the faulty exon harboring a frameshift mutation, using small pieces of DNA (antisense oligonucleotides). Bypassing the frameshift mutation facilitates reading of the remaining exons. Animal work confirmed the principal in mouse models and now clinical trials have started in Duchenne muscular dystrophy and Huntington disease.

Strategies to reduce the formation of toxic metabolites

Reduction of metabolic load

For IMDs in which the pathological consequences are related to the accumulation of toxic substances, one of the simplest and most broadly used therapeutic strategies is to reduce the load on the block in the pathway, by reducing the amount of substrate to be processed by that enzyme. Phenylketonuria is the classic example: blood phenylalanine concentrations are reduced by limiting phenylalanine, by strict restriction of normal dietary protein intake. Phenylalanine is an essential amino acid and must not be removed from the diet completely so a small amount of natural protein is eaten to supply phenylalanine requirements. Measurement of blood phenylalanine concentrations is performed to monitor dietary control, usually by taking fingerprick samples onto Guthrie cards, which are sent to the laboratory by the family. Phenylalanine free amino acid supplements are prescribed to prevent dietary deficiency from the severe protein restriction. The availability of artificial low-phenylalanine foods on prescription has greatly improved the tolerability of the diet, owing to their variety and greater acceptability to the patient.

Reduction in metabolic load is also used acutely in a number of conditions that present with decompensations. In urea cycle defects, protein intake is stopped during the initial presentation, while other treatment strategies that help reduce the hyperammonaemia take effect. A similar approach is used in organic acidaemias, to reduce organic acid production. Once recovery is apparent, protein is re-introduced and intake is gradually increased. Removal of protein for longer than 72 h risks exacerbating catabolism, including that of muscle protein, with further production of organic acids (or of ammonia in urea cycle defects). It is therefore usual practice to introduce some protein after 48–72 h to promote anabolism. Patients with the severest blocks will still be able to tolerate 0.5 g/kg/24 h of protein for synthetic requirements and repair, without precipitating further decompensation.

Blockage of formation of toxic metabolites

In tyrosinaemia type I, the deficiency of fumarylacetoacetase results in the accumulation of fumarylacetoacetate, which causes hepatorenal toxicity. Nitisinone (NTBC) is a drug used to prevent the production of these toxic metabolites, by blocking the catabolic pathway of tyrosine at an earlier stage (4-hydroxyphenylpyruvate dioxygenase) (Fig. 24.10). This means that the patient has, in effect, been converted from having the severe, tyrosinaemia type I, phenotype to the milder, tyrosinaemia type III, phenotype. Clinically, tyrosinaemia type I presents with acute metabolic decompensation, associated with severe liver compromise and renal impairment and, in the longer term, the development of hepatocellular carcinoma. The clinical picture in type III is much milder, with some neurological involvement reported in some patients: liver and renal disease are not features. In view of the continued block in tyrosine catabolism, albeit at an earlier stage, a low tyrosine diet is still required, but the need for liver

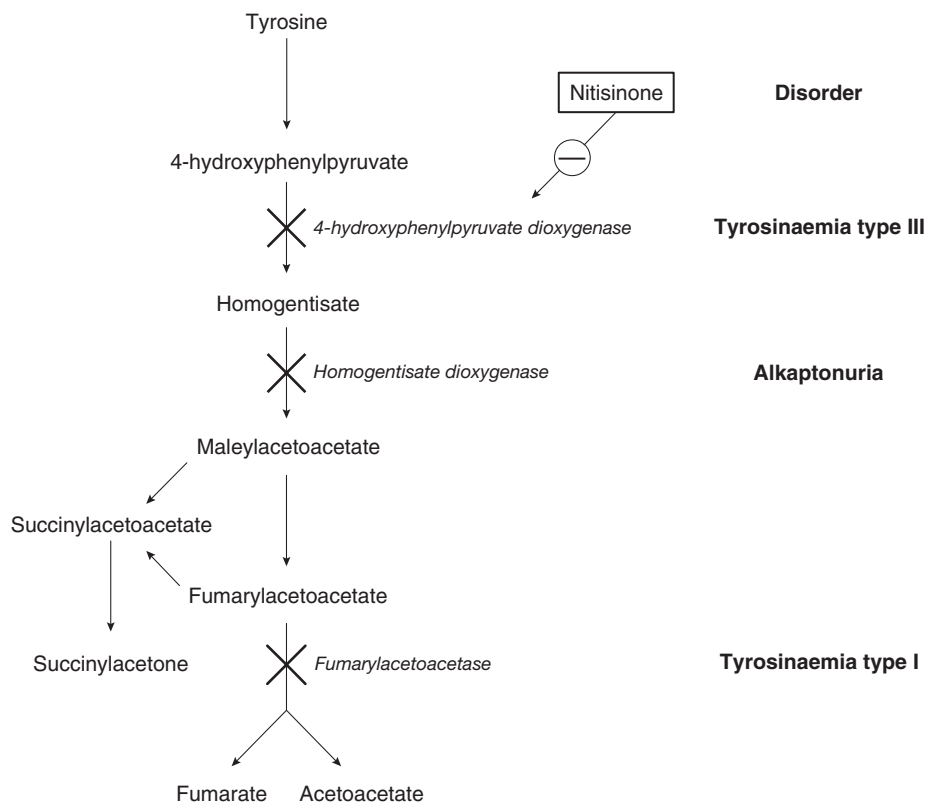


FIGURE 24.10 ■ Blocking the production of toxic metabolites: nitisinone blocks tyrosine catabolism at the level of 4-hydroxyphenylpyruvate dioxygenase, thus preventing the production of the toxic metabolites maleylacetoacetate and fumarylacetoacetate; this modifies the clinical phenotype to that of the milder tyrosinaemia type III from the more severe type I.

transplantation for hepatocellular carcinoma is dramatically reduced. Paradoxically, patients with milder tyrosinaemia type I that presents later, now do less well than those that present as neonates, as more hepatic fibrosis occurs before presentation and the introduction of NTBC.

Similarly, tyrosinaemia type III has a milder phenotype than alkaptonuria, caused by deficiency of homogentisate dioxygenase, the third enzyme in the pathway (see Fig. 24.10). This recessive condition, one of the conditions on which Garrod based his inherited metabolic traits theory, is manifest clinically by darkening of the urine on standing (usually presenting as staining of nappies). In the longer term, darkening of the sclerae and ear cartilage is observed, with subsequent darkening of the skin, especially over the nose, cheeks and axillae. In alkaptonuria, the joint involvement leads to arthritis, which can be severely debilitating. Low dose NTBC is now being evaluated in the management of this disorder.

Blockage of site of action of toxic metabolites

The effects of toxic metabolites can be limited by blocking their site of action. An example of this strategy is the use of dextromethorphan in the management of non-ketotic hyperglycinaemia, a glycine cleavage defect which results in high plasma concentrations of glycine. Glycine is a neurotransmitter with a central excitatory effect and inhibitory peripheral effect. Excessive concentrations, therefore, result in seizures with hypotonia. Infants

classically present with increased fetal movements (seizures in utero), hiccups and progressive apnoeas in the first days after birth, before developing encephalopathy and profound global developmental delay in the longer term. Sodium benzoate is used to reduce glycine levels, but this has little effect on long-term outcome. The use of partial antagonists of the N-methyl-D-aspartate (NMDA) receptor, the site of glycine binding, e.g. dextromethorphan and ketamine, improve seizure control, but do little to improve developmental outcomes.

Strategies to remove toxic substances

It may not be possible to reduce toxin production, and in acute decompensation, rapid clearance of toxic metabolites is required. This can be achieved either with the use of drugs that permit alternate pathways of excretion, or with dialysis or haemofiltration.

Drugs

Carnitine therapy is widely used in the management of IMDs for conjugating toxic metabolites and facilitating their urinary clearance as carnitine conjugates. This therapeutic approach is particularly used in organic acidaemias, e.g. propionic acidaemia and glutaric aciduria type I. Glycine is used in a similar way in isovaleric acidaemia. In urea cycle defects, sodium benzoate and sodium phenylbutyrate are used to conjugate with glycine and glutamine, respectively, which can then be excreted

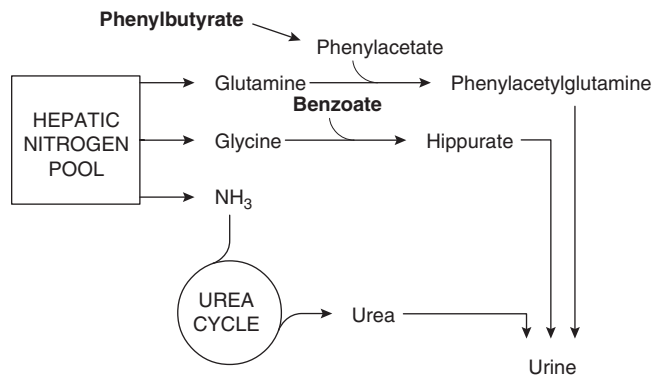


FIGURE 24.11 ■ Alternate pathway therapy in urea cycle defects; phenylbutyrate is metabolized to phenylacetate, which can combine with glutamine, allowing the latter to be excreted as phenylacetylglutamine; benzoate can combine with glycine to form hippurate, which can be excreted in the urine.

in the urine, bypassing the urea cycle and reducing its metabolic load (Fig. 24.11). However, as conjugation may be immature in neonates, these drugs may prove less effective in the newborn period. In organic acidemias, this treatment strategy is used to help control hyperammonaemia, which may cause sufficient vomiting to perpetuate the catabolic state driving metabolic decompensation.

Dialysis and haemofiltration

These techniques can be used effectively in acute settings, to clear toxic metabolites from the bloodstream. Originally, exchange transfusion was used for this purpose, but this procedure is inefficient and not without complications. Continuous venovenous haemofiltration, although not as efficient as haemodialysis at clearing toxic metabolites, has less impact on haemodynamic status. It is very effective at clearing lactate, ammonia, leucine and organic acids. Peritoneal dialysis is used in infants <2.5 kg, owing to the technical difficulties of performing haemodialysis on small babies.

Additional treatments

Appropriate nutritional support is essential to promote anabolism and prevent the catabolic drive to decompensation. To ensure that adequate calories are provided, when patients with IMDs require intravenous infusions, they are given fluids containing higher concentrations of glucose than are used in intravenous fluids administered to most other patients. A minimum of 10% dextrose (with appropriate electrolyte additives) is used in all age groups. Patients with congenital lactic acidosis are an exception as, in this group, the use of high concentrations of dextrose may exacerbate hyperlactataemia; in these patients, 5% dextrose-based solutions are used. Fat emulsions are used to provide additional calories, provided that a fat oxidation defect has been excluded. Patients known to have IMDs, are instructed to use an emergency regimen to help to reduce catabolism, during episodes of intercurrent

infection. This comprises a glucose polymer drink that has the advantage of easy absorption without a great osmotic load, thus reducing the risk of diarrhoea. These drinks can be administered at home during intercurrent illnesses, but if they are not tolerated, owing to vomiting or diarrhoea, admission to hospital is required, for intravenous dextrose therapy.

Substrate depletion

Different strategies are needed to manage storage disorders. Cystinosis is an autosomal recessive disorder presenting with a severe renal Fanconi syndrome and growth failure. The biochemical basis is a defective lysosomal membrane cystine transporter, leading to intracellular cystine accumulation. Cysteamine has been used successfully to reduce the progression of renal glomerular damage and improve growth. Cysteamine can combine with cystine to produce cysteine and a mixed disulphide cysteine-cysteamine, which can be exported by the cysteine and the lysine transporters, respectively.

Substrate deprivation

Substrate deprivation therapy is used in the management of lysosomal storage disorders. The principle is to reduce the amount of substrate synthesized so that it matches the impaired rate of catabolism. In this way, substrate does not accumulate. Miglustat is an inhibitor of glucosylceramide synthase, an essential enzyme in glycosphingolipid synthesis. It is used in the management of Gaucher disease type I and Niemann Pick C and is being investigated for the management of late onset Tay–Sachs disease.

CONCLUSION

Although the IMDs are individually rare, collectively they are an important cause of morbidity and mortality, particularly in the newborn, but in some cases in later life. Although the presentation of some IMDs is characteristic, for many it is non-specific, for example with growth failure, acidosis or hypoglycaemia. Nevertheless, the systematic use of investigations, starting with the simple and proceeding to more complex, will usually allow a diagnosis to be established and, when available, treatment to be commenced. Treatment is usually directed at overcoming the effect of the defective enzyme, but strategies for enzyme replacement are being developed and gene therapy may become more widely available in the future.

Further reading

Green A, Morgan I, Gray J. *Neonatology and laboratory medicine*. London: ACB Venture Publications; 2003.

This book covers acquired, as well as inherited, neonatal illness, with chapters on presenting features (e.g. jaundice) and prenatal diagnosis as well as a summary chapter on inherited metabolic disorders.

Scriver CR, Beaudet AL, Sly WS et al. editors. *The metabolic and molecular bases of inherited disease*. 8th ed., New York: McGraw-Hill; 2001. www.OMMBID.com.

The ultimate reference work for inherited metabolic diseases.

Paediatric clinical biochemistry

Fiona Carragher

CHAPTER OUTLINE

INTRODUCTION 484

POSTNATAL INVESTIGATION OF THE SMALL FOR GESTATIONAL AGE NEWBORN 484

Intrauterine infections 484

Maternal drug abuse 485

RESPIRATORY DISORDERS 485

Respiratory distress 485

Apnoea of prematurity 486

RENAL FUNCTION 486

Hyponatraemia 487

Hypernatraemia 487

Hydrogen ions 488

Interpretation of renal function tests 488

CARBOHYDRATE METABOLISM 488

Neonatal hypoglycaemia 489

CALCIUM AND PHOSPHORUS METABOLISM 490

Disorders of calcium and phosphorus metabolism 490

DISORDERS OF LIVER FUNCTION 493

Bilirubin metabolism 493

Unconjugated hyperbilirubinaemia: physiological jaundice 493

Unconjugated hyperbilirubinaemia: pathological causes 494

Conjugated hyperbilirubinaemia and hepatocellular disease 494

Liver disease in older children 495

INTRODUCTION

National statistics for the UK record approximately 700 000 live births per year. In England and Wales in 2010, the infant mortality rate was 4.3 deaths per 1000 live births, the lowest ever recorded, which compares with an infant mortality rate of 12 deaths per 1000 live births in 1980. This considerable decrease in infant mortality can be partly explained by improvements in health care and more specifically, in midwifery and neonatal intensive care. The neonatal period, defined as the first four weeks of life, is associated with the highest mortality rate in childhood, with infants of very low (<1500 g) and low (<2000 g) birth weight accounting for a significant proportion of deaths. The mortality rate reflects the consequences of congenital malformation and disease and the difficulties of transition from intra- to extrauterine life, particularly in premature infants; such a high rate is not seen again until the seventh decade of life. For these reasons, the majority of this chapter is devoted to neonatal clinical biochemistry. Disorders of older children are only included where there are particular problems with interpretation of clinical biochemistry data.

The most common cause of admission to a neonatal intensive or special care unit is prematurity. 'Pre-term' in the practical sense rarely includes neonates of more than 32 weeks of gestation at birth. The problems of prematurity can be related to immaturity of vital systems, for example the lungs, kidneys and central nervous system, or inadequacy of energy stores or of nutrients such as

calcium and iron that are accreted during the third trimester of pregnancy. Another high-risk group is those neonates with a birth weight inappropriate for their gestational age. In most cases, these babies are small for gestational age (SGA), defined as birth weight below the 10th centile for newborns of the same gestational age in the same population. However, neonates born to mothers with diabetes may be overweight as a result of hyperglycaemia in utero.

POSTNATAL INVESTIGATION OF THE SMALL FOR GESTATIONAL AGE NEWBORN

Inappropriate birth weight is often readily explicable from the clinical history. Common causes of growth retardation are multiparity, pre-eclampsia, infection and drug addiction. An uncommon example would be a first-born SGA infant with microcephaly, the mother of whom should be investigated for possible hyperphenylalaninaemia by analysis of maternal plasma amino acids.

Intrauterine infections

Antenatal infections can affect the developing fetus, leading to congenital malformations and/or low birth weight. Many viruses have teratogenic effects on the fetus, so maternal infection in the first trimester, when organogenesis is occurring, tends to cause the greatest effect. In the UK, universal

TABLE 25.1 Reference ranges (5–95th centile) for serum immunoglobulins (g/L) in UK Caucasians

Age	IgG	IgA	IgM
Term cord blood	5.2–18.0	<0.02	0.02–0.20
0–2 weeks	5.0–17.0	0.01–0.08	0.05–0.20
2–6 weeks	3.9–13.0	0.02–0.15	0.08–0.40
6–12 weeks	2.1–7.7	0.05–0.40	0.15–0.70
Adult	6.0–16.0	0.80–2.80	0.50–1.90
15–45 years			

From the PRU Handbook of Clinical Immunochemistry – see Milford-Ward et al. in Further reading.

screening for syphilis, hepatitis B, HIV and susceptibility to rubella infection is offered to all mothers in the early stages of pregnancy, with follow-up diagnosis and treatment available for both mother and infant as appropriate.

Intrauterine infection is indicated by an increase in the infant's serum IgM concentration (Table 25.1) and is often associated with prolonged unconjugated hyperbilirubinaemia and abnormal liver function tests. Measurement of the serum IgM concentration is only useful as an indicator of intrauterine infection before six weeks of age. After this time, the normal rise in serum IgM concentration may mask any early increase, and investigation requires specific serological testing.

Maternal drug abuse

Two percent of women in urban areas of the UK have been reported to have had a positive screening test for at least one prohibited substance at the time they first attended for antenatal care. These positive tests included amphetamines, barbiturates, cannabinoids, cocaine metabolites, methadone and opiates.

The consequence of maternal drug abuse is dependent upon the intensity of use. Infants may be born prematurely and/or SGA, and a proportion exhibit withdrawal symptoms during the neonatal period. The appearance of these symptoms depends upon the rates of clearance of the drugs and their active metabolites. For example, the irritability, tremors and convulsions of heroin withdrawal appear during the first 24h of life, whereas the same symptoms do not generally appear until after 24–48h when methadone is the substance involved.

A urine drug screen of a symptomatic neonate is often negative. Poor drug penetration across the placenta means that there are higher concentrations in maternal tissues and, consequently, it is more rewarding to screen maternal urine. The likelihood of detecting maternal drug abuse by investigating the urine of the neonate follows a decreasing order of cocaine > methadone > heroin > benzodiazepines.

RESPIRATORY DISORDERS

Respiratory distress

Up to 55% of the fetal cardiac output goes to the placenta. At birth there is an increase in peripheral resistance and a reduction in pulmonary vascular resistance, both of which

ensure the closure of the foramen ovale and constriction of the ductus arteriosus. Blood is then diverted through the pulmonary vessels and the adult type of circulation is established. In newborn infants, haemoglobin F (HbF) is the predominant haemoglobin, accounting for about 75% of the total. Haemoglobin F has a greater affinity for oxygen than haemoglobin A (HbA) and its oxygen dissociation curve is shifted to the left, ensuring adequate oxygen exchange at a significantly lower partial pressure of oxygen (PO_2). As the partial pressure of arterial oxygen (PaO_2) in blood rises with postnatal age, blood HbF concentrations fall, and by six months of age HbF accounts for only about 5% of the total haemoglobin.

Respiratory distress syndrome (RDS) may be caused by a variety of conditions (see Box 25.1). The most common cause in pre-term infants is hyaline membrane disease, primarily attributable to immature surfactant synthesis. Surfactant ensures patency of the alveoli by reducing the surface tension of the alveolar wall, and is made up of phosphatidylcholine (lecithin), phosphatidylglycerol and phosphatidylinositol. Surfactant synthesis begins by the 20th week of gestation. It increases slowly until the 34th week and then more rapidly as the type II alveolar cells mature. The rate of synthesis is sensitive to cold, hypoxia and acidemia and may be halved by postnatal exposure to temperatures <35°C or an arterial $[H^+] > 56$ nmol/L (pH <7.25). High intrauterine glucose concentrations, resulting from poorly controlled maternal diabetes mellitus, can also delay maturation of fetal surfactant synthesis. Fetal type II alveolar cell maturation may be enhanced by maternal steroid therapy. Corticosteroids are often given to women who are at risk of premature delivery at 24–34 weeks of gestation.

The incidence of hyaline membrane disease is inversely related to postconceptional age. The signs of the disease, which occur within four hours of birth, include sternal retraction, intercostal and subcostal recession, expiratory grunt and tachypnoea with a respiratory rate >60 per min. A characteristic 'ground glass' appearance is present on radiological examination of the chest. The appearance

BOX 25.1 Some causes of neonatal respiratory distress

Pulmonary

- Hyaline membrane disease
- Pneumonia
- Meconium aspiration
- Transient tachypnoea of the newborn
- Pneumothorax
- Pulmonary haemorrhage

Non-pulmonary

- Patent ductus arteriosus
- Acute blood loss
- Congenital heart disease
- Hypo/hyperthermia
- Metabolic acidosis
- Polycythaemia
- Intracranial birth trauma

of these reticulogranular opacities allows diagnosis of hyaline membrane disease with 90% confidence. Other entities that may produce similar opacities include immature lung, wet lung disease, neonatal pneumonia, idiopathic hypoglycaemia, congestive heart disease, maternal diabetes and early pulmonary haemorrhage.

Group B streptococcal pneumonia may present as early as four hours after infection, during birth, from this common vaginal organism. It occurs more often in the pre-term than the full-term infant. Measurement of serum C-reactive protein and microbiological investigations can be used to distinguish bacterial pneumonia from hyaline membrane disease.

Meconium aspiration and transient tachypnoea of the newborn (TTN) occur most often in full- or post-term infants. The passage of meconium (the contents of the fetal bowel) in utero is associated with fetal hypoxia. In some cases, the contaminated liquor is aspirated. Clinical symptoms usually appear 12–24 h after birth. Transient tachypnoea of the newborn occurs in term neonates, usually those born by caesarean section, and appears to be due to incomplete stimulation of adrenergic mechanisms for lung clearance during birth. Pneumothorax may occur as a complication of either of these conditions or as a result of mechanical ventilation of pre-term infants.

In RDS, the neonate develops hypoxia and a respiratory acidosis, both of which cause an increase in pulmonary vascular resistance and thus pulmonary hypertension with a left-to-right shunt. Hypoxia enhances anaerobic glycolysis and may result in lactic acidosis. Additional complications include brain damage and oedema, and hypotension that may lead to renal failure, paralytic ileus and/or necrotizing enterocolitis (NEC).

Non-pulmonary causes of respiratory distress are usually self-evident and normally improve with treatment of the underlying conditions (see Box 25.1). Up to 20% of infants weighing <1750 g at birth have a patent ductus arteriosus (PDA). Medical management of PDA includes fluid restriction and stimulation of diuresis. Pharmacological closure may be achieved using indometacin, an inhibitor of prostaglandin synthetase. Contraindications to this treatment include renal insufficiency (plasma creatinine concentration >160 µmol/L) with or without oliguria, bleeding disorders and NEC.

Management of respiratory distress

This may involve assisted ventilation with oxygen, aiming to maintain the P_{aO_2} in the range 6–12 kPa, P_{aCO_2} 5.5–8.0 kPa and arterial O_2 saturation (SaO_2) 88–95%. Although careful monitoring of blood gases is required, repeated blood sampling may cause anaemia. Anaemia may necessitate transfusion, usually with adult haemoglobin. Increased HbA may compromise oxygen uptake in the lungs in the presence of low alveolar P_{O_2} , thereby further aggravating tissue hypoxia. Correction of the metabolic acidosis with sodium bicarbonate can cause oedema and precipitate heart failure, owing to sodium and water overload. Too much oxygen may result in retrolental fibroplasia with retinal detachment and blindness.

Long-term ventilation is associated with bronchopulmonary dysplasia, hyperinflated emphysematous lungs with extensive alveolar destruction and widespread fibrosis. The proposed role of free oxygen radicals in the development of these complications has led many paediatricians to administer vitamin E routinely to all infants on ventilators.

Transcutaneous PO_2 (TcO_2) polarographic electrodes can be used to monitor oxygen treatment in infants who have good skin perfusion; the results correlate well with arterial PO_2 measurements. Transcutaneous PO_2 is measured at a skin temperature of 44 °C so the electrodes require frequent re-siting with recalibration in order to prevent skin burns. These problems do not occur with pulse oximetry, which measures SaO_2 of oxyhaemoglobin and reduced haemoglobin during an arterial pulse, by differential light absorption at 660 and 940 nm. The results correlate well with arterial PaO_2 measurements at SaO_2 values >65%, but are falsely low in the presence of methaemoglobin because its molar extinction characteristics imitate reduced haemoglobin.

Surfactant administered through an endotracheal tube significantly reduces the incidence and complications of respiratory distress in newborns at risk of developing hyaline membrane disease.

Apnoea of prematurity

Apnoea of prematurity, defined as a cessation of breathing for >20 s, with or without bradycardia and cyanosis, occurs in up to 85% of infants weighing <1000 g at birth. The major cause of apnoea is immaturity of the central respiratory drive, with poor sensitivity to changes in $PaCO_2$. This is compounded by the suppressed respiratory response to hypoxia, which serves to reduce oxygen requirements in utero and which persists in the pre-term infant. Poor coordination of the major respiratory muscles of the chest wall and the upper airways can lead to obstruction, usually at the level of the pharynx. Respiratory effort against a closed airway distorts the chest wall and activates the intercostal phrenic inhibitory reflex. Apnoea is worsened by infection, thermal instability and hypoglycaemia.

Apnoea of prematurity is treated with methylxanthines, which augment central respiratory drive and increase the sensitivity of chemoreceptors to changes in $PaCO_2$.

RENAL FUNCTION

The investigation of renal disorders and the monitoring of fluid and electrolyte replacement in pre-term infants are complicated by immaturity of organ function, and by the difficulty in collecting accurately timed urine samples.

Nephrons develop from about the sixth week of gestation and start producing urine from about the tenth week. The full complement of nephrons is present by the 36th week of gestation. Glomerular function develops more rapidly than tubular function. At term, the tubules are relatively short and underdeveloped. They increase in length and develop increasing absorptive and secretory

TABLE 25.2 Daily fluid and electrolyte requirements in infants and children compared with adults

	Water (mL/kg)	Sodium (mmol/kg)	Potassium (mmol/kg)
<1 month	150	2–6 (mean 4)	2–6 (mean 4)
<1 year	100		
1–3 year	90		
3–6 year	80		
6–9 year	70		
9–12 year	60		
Adult	30–40	1.5	1

function during the neonatal period. Functional immaturity is characterized by an inappropriately high urinary sodium excretion for the glomerular filtration rate (GFR) and an impaired response to a sodium load. This is pertinent to the management of pre-term infants, in whom the sodium requirement per kg body weight may be higher than that for term or older infants, with more mature tubular function (Table 25.2). The loops of Henle are principally juxtamedullary in position: they, too, are relatively short compared with those in older infants and adults and do not penetrate deeply into the renal medulla, thereby limiting renal concentrating ability.

In low birth weight infants, glomerular function is adequate for growth but may be inadequate to cope with the increased nitrogenous load during periods of catabolism, starvation, hypoxia, infection and infusion of nitrogen-containing solutions. Tubular function is adequate for the filtered load associated with a reduced GFR and can normally maintain an appropriate excretory function and systemic acid–base status during the anabolic growth phase.

The GFR increases with postconceptional age as renal blood flow increases and renal vascular resistance decreases; full functional maturity is not reached until about the second year of life.

Total body water constitutes about 85% of the body weight of pre-term infants weighing <1.0 kg at birth, compared with about 75% in term infants and 60% in adults (Fig. 25.1). Relatively more water is in the extracellular compartment than in the intracellular compartment. Total body sodium is about 45 mmol and potassium 75 mmol, and blood volume about 70 mL. During the first few weeks of life, there is a contraction of the extracellular space associated with an increase in urinary sodium loss. This results in an initial 10–15% reduction in body weight. The onset of natriuresis coincides with improvement of lung function and is probably related to a reduction in pulmonary vascular resistance and the release of natriuretic peptides. This contributes to the contraction of the extracellular fluid space. Excess fluid and sodium administration, before the onset of the natriuretic phase, may worsen respiratory distress, delay the closure of the ductus arteriosus and precipitate oedema with hyponatraemia. It is not appropriate to rely on spot urinary sodium concentrations to monitor water and sodium requirements during the first few weeks of life in a premature infant.

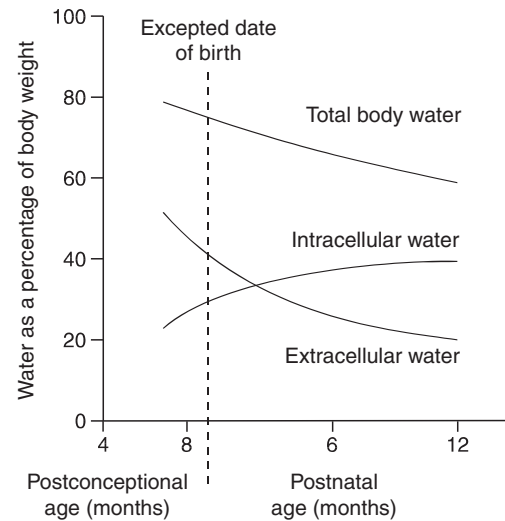


FIGURE 25.1 ■ Changes in body water as a function of age for pre-term and term infants. The timescale indicates expected date of birth and postnatal age.

Antidiuretic hormone (ADH) secretion, in response to volume reduction and hyperosmolality, occurs from about the 25th week of gestation. The interstitial osmolality of the renal medulla, which is the main determinant of the urine concentrating ability in the presence of ADH, is dependent on the countercurrent multiplication mechanism in the loops of Henle and on the interstitial concentrations of sodium, chloride and urea. The reduced GFR and urea clearance decrease the tubular reabsorption of urea and the interstitial urea concentration; consequently, the concentrating ability is impaired, despite appropriate ADH output. Even full-term infants have a limited capacity to conserve water, and urine osmolality rarely rises >700 mmol/kg. Approximate fluid requirements are shown in Table 25.2.

Hyponatraemia

In neonates, hyponatraemia may be caused by maternal fluid retention or overload during labour, or excess neonatal hypotonic fluid administration during the postnatal period. In addition, inappropriate antidiuresis may develop as a consequence of respiratory disease or intraventricular haemorrhage. In older infants, a dilutional hyponatraemia may be seen, caused by a combination of water retention and sodium depletion in response to increased intestinal or renal fluid losses. Signs of hyponatraemia are related to the rate of fall of plasma sodium concentration rather than to its actual value, and may include hypotension, drowsiness and convulsions. Congenital adrenal hyperplasia (CAH) must always be considered as a possible cause of hyponatraemia (see Chapter 21).

Hypernatraemia

As in adult patients, hypernatraemia may be caused either by water depletion or by excess sodium administration or

retention. Insensible water loss is significantly greater in pre-term infants compared with children and adults. The reasons include:

- greater surface area to body volume ratio
- increased skin blood flow
- increased metabolic and respiratory rates
- lack of subcutaneous fat
- greater transepidermal fluid loss.

The epidermis of the skin matures by about the 28th week of gestation with keratinization of the stratum corneum. Consequently, infants born before 28 weeks are at greater risk of excess fluid loss and may lose up to 60 mL/kg/24h of water through the skin, compared with about 10 mL/kg/24h in term infants. Additional environmental factors, such as temperature and humidity, also affect transepidermal fluid loss.

Impaired response to changes in blood volume and plasma osmolality make infants particularly vulnerable to developing hypernatraemia. Increased urinary free water loss may be caused by glycosuria secondary to either hyperglycaemia or to the low renal threshold for glucose. In pre-term infants, glycosuria may be present when the plasma glucose concentration is as low as 5.6 mmol/L. Diabetes insipidus, caused, for example, by intracranial injury at birth, may also result in an increase in free water clearance.

Although the major cause of sodium and water imbalance in premature infants is related to the functional immaturity of organs, it is important to be aware that some drugs may exacerbate the situation, e.g. sodium bicarbonate, indometacin and methylxanthine derivatives such as caffeine.

Morbidity and mortality from hypernatraemia are caused by the increased extracellular osmolality and its effects on fluid distribution between fluid compartments. Infants with hypernatraemia may present clinically with irritability and lethargy, convulsions and coma; lesions within the CNS include intracerebral and intraventricular haemorrhages, as well as sinus and small vessel thromboses. The majority of infants who survive suffer no long-term sequelae, although a minority develop persistent neurological abnormalities.

Hydrogen ions

Under normal circumstances, the kidneys can excrete the hydrogen ion load and generate sufficient buffering capacity to maintain normal systemic acid–base status, despite apparent immaturity of renal tubular function. However, if hydrogen ion production is increased, premature infants are prone to develop a metabolic acidosis in addition to the respiratory acidosis of respiratory distress. The proximal tubular threshold for bicarbonate reclamation is reduced, as is the distal tubular response to an ammonium chloride load: thus the generation of buffering capacity is reduced. In addition, urinary phosphate excretion, which is dependent on GFR, phosphorus intake and plasma phosphate concentration, may be significantly reduced in infants with phosphate depletion.

Metabolic acidosis is frequently related to disorders of respiratory and cardiac function. Echocardiography is invaluable in the investigation of congenital abnormalities of the heart and of left ventricular function. Inherited metabolic diseases, such as mitochondrial dysfunction with lactic acidosis and organic acidaemias, should be excluded early by appropriate investigations (see Chapter 24).

Interpretation of renal function tests

Renal function tests must be interpreted with caution in pre-term infants. Despite the low glomerular filtration rate (GFR), plasma urea concentrations are low in neonates compared with adults, because of increased utilization of nitrogen. Plasma concentrations fluctuate markedly in response to changes in the catabolic rate, GFR and nitrogen intake. Plasma creatinine concentrations are high at birth, being determined principally by maternal concentration: they fall initially and then gradually increase with respect to postnatal age and muscle mass. Measurements may be subject to analytical variation owing to negative interference by bilirubin, and reference intervals may therefore be method dependent.

The interpretation of tubular function tests may be complicated in premature infants in whom the presence of a generalized aminoaciduria and glycosuria may be misinterpreted as a type 2 renal tubular acidosis (RTA). However, despite immature tubular function, maximum phosphate reabsorption occurs in premature infants when plasma phosphate concentrations fall below about 1.5 mmol/L. Impaired response to an ammonium chloride load in the presence of a systemic acidosis may be wrongly interpreted as a type 1 RTA, rather than as a result of immaturity of distal tubular function.

CARBOHYDRATE METABOLISM

The fetus receives a constant, transplacental supply of glucose and amino acids, maintained by maternal intermediary metabolism. Although free fatty acids are transported across the placenta, they are not used for oxidative metabolism but are stored in adipose tissue as triglycerides (triacylglycerols). At birth, the umbilical vein plasma glucose concentration is about 70% that of the maternal circulation. After birth, the plasma glucose concentration falls, resulting in an increase in the circulating concentrations of glucagon, catecholamines and growth hormone (GH), which, together with a fall in plasma insulin concentration, stimulate the release of glucose from hepatic glycogen and activate lipolysis and gluconeogenesis (Fig. 25.2). Increased hepatic cAMP synthesis is associated with glycogenolysis and, through induction of synthesis of phosphoenolpyruvate carboxykinase, gluconeogenesis. Substrates such as pyruvate, lactate and alanine, which are diverted to gluconeogenesis, are no longer available to form malonyl-CoA and hence fatty acids. As the cytosolic concentration of malonyl-CoA falls, its inhibitory effect on carnitine palmitoyltransferase I activity is

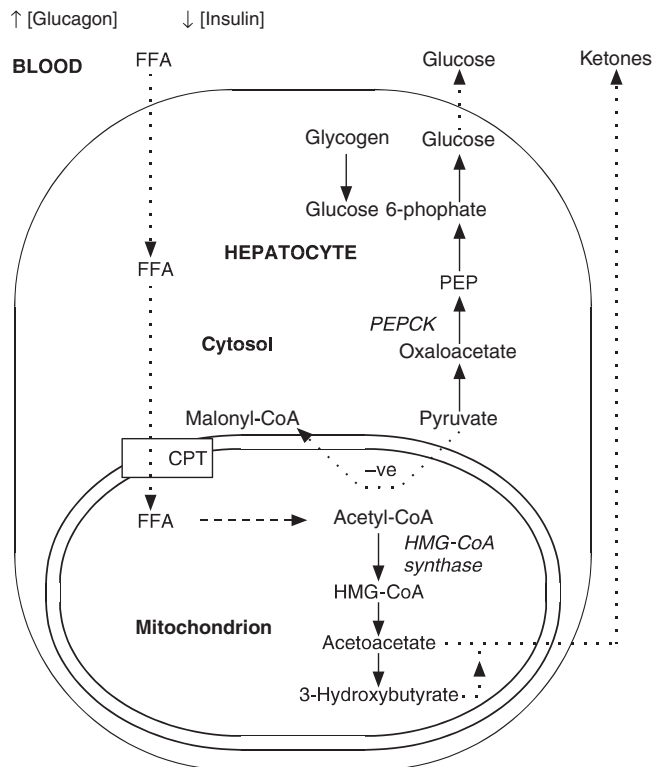


FIGURE 25.2 ■ Changes in energy metabolism at birth. Plasma glucagon rises and insulin falls after birth, in response to falling plasma glucose concentrations. Hepatic glycogenolysis is stimulated. Increased synthesis of phosphoenolpyruvate carboxykinase (PEPCK) stimulates gluconeogenesis. The -ve sign indicates that pyruvate is no longer used to form malonyl-CoA. Malonyl-CoA-dependent inhibition of carnitine palmitoyltransferase (CPT) is reduced and free fatty acids (FFA) can be transferred into the mitochondria where they are subject to β -oxidation. Activation of 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) synthase permits hepatic ketone synthesis. The dotted lines indicate transfer (active or passive) across membranes.

reduced. Carnitine palmitoyltransferase I plays a role in transferring long chain free fatty acids, released from triglycerides through the action of hormone-sensitive lipase, into the mitochondria where they undergo β -oxidation. Activation of 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) synthase enhances the hepatic formation of ketones from acetyl-CoA.

Although glucose is the major source of energy for the brain, fetal and neonatal brains have an enhanced capacity to utilize ketones, lactate and amino acids as alternative fuels. Circulating free fatty acids and glycerol appear within 3 h after birth in the fasting full-term infant, but plasma concentrations of ketones begin to rise only after 12–48 h of fasting. This relative resistance to ketonaemia is caused by the high turnover of ketones in the neonate, for example ketones can be metabolized at a rate of 13–22 pmol/kg/min, after only 8 h of fasting and may account for up to 25% of the infant's basal energy requirements. Such a high ketone turnover is observed only after several days of fasting in adults.

The maintenance of normal plasma glucose concentrations during the immediate postpartum period

depends upon regulated insulin and glucagon secretion, adequate reserves of both glycogen and triglycerides and the possession of operational intermediary metabolism. When all these factors are present and functioning appropriately, glucose production can reach 5–8 mg/kg/min in full-term infants, compared with about 2 mg/kg/min in adults. This requirement for a high rate of glucose production is related to the higher brain to body mass ratio of infants.

Neonatal hypoglycaemia

Hypoglycaemia of the newborn is defined as a blood glucose concentration of <2.5 mmol/L. It occurs in up to 0.4% of all newborns, but is much more common in infants of diabetic mothers (previously diagnosed as having diabetes or with gestational diabetes) and in low birth weight and premature infants. Signs of hypoglycaemia include:

- jitteriness
- feeding problems
- abnormal cry
- floppiness
- pallor
- apnoea
- cyanosis
- convulsions.

The severity of symptoms will, in part, depend on the rate of fall of the plasma glucose concentration and on the infant's ability to adapt to alternative sources of energy. In all cases, early diagnosis and the prompt restoration of normoglycaemia are critical in order to minimize damage to the developing CNS.

Infants at risk of developing hypoglycaemia should be monitored regularly. Hypoglycaemia detected by point-of-care testing should be confirmed by laboratory testing, as falsely low results may be obtained in the presence of a high packed cell volume and poor peripheral perfusion of the site of testing.

The aetiology of neonatal hypoglycaemia must be identified so that causes of persistent hypoglycaemia associated with rare disorders and inherited diseases of intermediary metabolism can be excluded (see below). Some causes of hypoglycaemia that may occur during the neonatal period or later in infancy are shown in [Box 25.2](#).

The major cause of hypoglycaemia in pre-term infants is inadequacy of energy stores that would normally have accumulated during the last trimester of pregnancy. The glycogen content of fetal liver increases by 300%, from 0.9 to >3.9 g/100 g, between 31 and 40 weeks of gestation. Small for gestational age infants also have reduced energy stores and they, too, are at risk of hypoglycaemia. Infants in shock, for example following perinatal asphyxia or because of neonatal sepsis, may develop hypoglycaemia owing to underperfusion of the liver.

Hyperinsulinism is the most common cause of severe and persistent hypoglycaemia in term infants. Hyperinsulinaemic infants are particularly prone to complications of hypoglycaemia, as the high insulin

BOX 25.2 Some causes of neonatal hypoglycaemia

- Inadequate energy stores
 - Prematurity
 - Small for gestational age
- Impaired hepatic glucose production
 - Perinatal asphyxia
 - Neonatal sepsis
 - Hormone deficiency (hypopituitarism, adrenal insufficiency)
- Transient hyperinsulinism
 - Infants of diabetic mothers
 - Erythroblastosis fetalis
 - Idiopathic transient hyperinsulinism
- Persistent hypoglycaemia
 - Persistent hyperinsulinaemic hyperglycaemia of infancy due to adenomatous or diffuse hyperplasia of the islets of Langerhans
 - Inherited metabolic diseases

concentrations inhibit lipolysis and the generation of alternative energy sources, such as ketones.

Hyperinsulinism in infants born to diabetic mothers is due to in utero hyperglycaemia-induced hyperplasia of the islets of Langerhans. Idiopathic transient hyperinsulinism may occur in interuterine growth-retarded infants and those suffering from perinatal asphyxia.

Persistent hyperinsulinaemic hypoglycaemia of infancy (PHHI), previously called familial hyperinsulinism, pancreatic nesidioblastosis or hyperinsulinaemic hypoglycaemia, causes recurrent severe hypoglycaemia that may present in term newborn infants or during the first few weeks of life. It can be subdivided into two types: focal adenomatous hyperplasia of the islets of Langerhans and diffuse (but discrete) hyperplasia. Mutations in several different genes that code for proteins involved with insulin release have been identified as being responsible. These include mutations associated with the two subunits of the β cell ATP-sensitive potassium channel – the sulphonylurea receptor proteins (SUR1) and the inward rectifier potassium channel proteins (Kir 6.2). Activating mutations in the glucokinase (GK) and glutamate dehydrogenase (GDH) genes have also been identified in this disorder. Glutamate dehydrogenase is a regulator of insulin secretion in pancreatic β cells and of ureagenesis in the liver, and defects in enzyme activity are characterized by hyperinsulinism with mild, asymptomatic hyperammonaemia. In many patients, the only therapeutic option is partial pancreatectomy, which carries the long-term risk of diabetes mellitus as apoptosis depletes the β -Cell mass. Those forms of PHHI resulting from mutations in GK and GDH are more amenable to pharmacological intervention, for example with diazoxide.

If the amount of glucose required to prevent recurrence of hypoglycaemia exceeds about 10 mg/kg/min, then hyperinsulinaemia is likely. If there is no known history of maternal diabetes mellitus or any of the other classic features (e.g. cherubic appearance and polycythaemia), blood samples taken during the hypoglycaemic episode should be assayed for insulin and 3-hydroxybutyrate.

In hyperinsulinaemic hypoglycaemia, the plasma insulin concentration is inappropriately high for the low plasma glucose concentration; plasma C-peptide concentrations are not a reliable reflection of pancreatic β -Cell function in neonates because of renal immaturity and reduced GFR. A plasma 3-hydroxybutyrate concentration <1.1 mmol/L during hypoglycaemia suggests that fatty acid oxidation and ketone synthesis are impaired. This can be due to hyperinsulinaemia, inadequate adipose stores or to a defect in β -oxidation of fatty acids.

The symptomatic treatment of hypoglycaemia is not without its complications, which include hyperglycaemia associated with hyperosmolality and cerebral hemorrhage, glycosuric osmotic diuresis and water overload.

CALCIUM AND PHOSPHORUS METABOLISM

Disorders of calcium and phosphorus metabolism are relatively common in pre-term infants. Calcium and phosphorus are actively transported across the placenta with fetal plasma concentrations higher than those in the maternal circulation. The transplacental transport is probably enhanced by locally synthesized 1,25-dihydroxy vitamin D ($1,25(\text{OH})_2\text{D}$) and by parathyroid hormone-related peptide (PTHrP), for which receptors have been identified in the placenta. The high plasma calcium concentration suppresses fetal parathyroid gland activity and probably inhibits fetal $1,25(\text{OH})_2\text{D}$ synthesis. During the last trimester of pregnancy, there is a three-fold increase in fetal calcium and phosphorus uptake (Table 25.3): the changes in fetal hormone concentrations facilitate bone mineralization and growth. Consequently, infants born prematurely during the last trimester of pregnancy are significantly disadvantaged and are prone to develop complications of calcium and phosphorus metabolism.

Disorders of calcium and phosphorus metabolism

Hypercalcaemia

Hypercalcaemia is relatively uncommon during the neonatal period. The condition may occur in phosphorus-depleted premature infants receiving unsupplemented breast milk or parenteral nutrition: it is invariably associated with hypophosphataemia. Treatment with phosphorus supplementation alone may precipitate acute hypocalcaemia. Idiopathic hypercalcaemia, with a normal or high

TABLE 25.3 Fetal calcium and phosphorus accumulation

Postconceptional age (weeks)	Body weight (kg)	Accretion rate (mmol/kg/day)	
		Calcium	Phosphorus
28	1.0	3.0	2.0
40 (term)	3.0	3.4	3.0

normal plasma phosphate concentration, is more common in mature infants but may occur in premature infants receiving high-dose vitamin D prophylaxis. It may also be associated with subcutaneous fat necrosis of infants. This disorder usually affects healthy full-term infants who have suffered perinatal asphyxia. These infants usually develop subcutaneous plaques over the buttocks and back. The condition is caused by excess production of $1,25(\text{OH})_2\text{D}$ by the granulomatous cells, similar to that observed in sarcoidosis in adults. Although very rare, it is the most common cause of hypercalcaemia in full-term neonates.

Hypercalcaemia during infancy and childhood is also rare but may be caused by immobilization, primary hyperparathyroidism, familial hypocalciuric hypercalcaemia or infantile hypercalcaemia (Williams syndrome). Familial hypocalciuric hypercalcaemia is a rare, autosomal dominant condition caused by mutations in the calcium sensing receptor (CASR). It is characterized by mild hypercalcaemia and a low urine calcium excretion and does not usually require treatment. Infants with mutations of the *CASR* gene may occasionally present in the neonatal period with severe hypercalcaemia (neonatal severe primary hyperparathyroidism). Features of Williams syndrome include failure to thrive, hypercalcaemia, elfin-like facial features, supravalvular aortic stenosis and learning disabilities. Microdeletion of the 7q11.23 region on chromosome 7, which contains the elastin gene, can be demonstrated by fluorescence in-situ hybridization (FISH).

Hypocalcaemia

Some causes of neonatal hypocalcaemia are shown in Box 25.3. After parturition, plasma total and ionized calcium concentrations fall, reaching a nadir between the 24th and 48th hour of life, then rising, such that the ionized calcium concentration is 1.10–1.40 mmol/L by

BOX 25.3 Some causes of neonatal hypocalcaemia

Early

- Prematurity (inversely related to postconceptional age)
- Infants with perinatal asphyxia
- Infants of diabetic mothers

Late

- Mineral depletion
 - Unsupplemented breast milk
 - Inadequate parenteral nutrition
 - Maternal vitamin D deficiency
- Exchange transfusion
- Iatrogenic
 - Diuretic therapy
 - High phosphorus intake
- Organ disease
 - Liver disease
 - Renal disease
- Endocrine disorders
 - Transient hypoparathyroidism
 - Hypoparathyroidism
 - DiGeorge syndrome

about the fifth day of life. Plasma ionized calcium concentrations as low as 0.70 mmol/L have been observed in pre-term infants without apparent clinical abnormalities. The degree of fall is inversely related to gestational age and is more marked in infants born to diabetic mothers and those with perinatal asphyxia. These changes, which are usually self-limiting, are probably caused by transient suppression of fetal parathyroid gland activity.

Permanent causes of hypocalcaemia include hypoparathyroidism, either X-linked or autosomal recessively inherited, or as part of the DiGeorge syndrome and pseudohypoparathyroidism. DiGeorge syndrome is associated with hypoplasia or aplasia of the parathyroid glands and congenital heart anomalies with partial or complete absence of the thymus. Many infants have a partial deletion of chromosome 22q11.

Pseudohypoparathyroidism is characterized by hypocalcaemia and hyperphosphataemia with appropriately increased secretion of parathyroid hormone (PTH) but end-organ resistance to its effect: in the common form of the condition this is caused by mutations in the *GNAS* gene. Clinically, infants with this disorder present with a constellation of features known as Albright hereditary osteodystrophy. Signs include short stature, dysmorphism, obesity and shortening of the fourth and fifth metacarpals. There may be associated hypothyroidism; subcutaneous calcification may also occur.

Osteopenia of prematurity

Significant failure of bone mineralization, or osteopenia, occurs in >50% of premature infants weighing <1000 g at birth. Osteopenia has a reported incidence of up to 13% in infants weighing <1500 g at birth.

Osteopenia of prematurity is caused by impaired mineralization of normal osteoid of the epiphyseal growth plate with proliferation of unmineralized osteoid. The clinical condition usually presents between the 6th and 12th postnatal weeks, associated with a fall-off of longitudinal growth, with continued growth in head circumference and frontal bossing. Swelling of the costochondral junctions of the ribs may result in the classic 'rachitic rosary'. Softening of the rib cage with impaired chest wall compliance and patchy lung atelectasis impairs ventilation and may cause late onset respiratory distress, so that ventilatory support may be required. Multiple non-traumatic fractures may also be present.

The aetiology of osteopenia of prematurity is multifactorial but the condition is primarily caused by deficient mineral substrate intake, particularly of phosphorus. Human breast milk is capable of providing up to 1.4 mmol/kg/24h of calcium and 0.9 mmol/kg/24h of phosphorus. Although these calcium and phosphorus intakes are adequate to enable normal bone mineralization and skeletal development in term infants, they fall far short of the mineral accretion that should have taken place during the last trimester of pregnancy and are not sufficient for a rapidly growing pre-term infant. Infants born as early as 26 weeks are able to hydroxylate vitamin D in the 25 position, but there may be maturational delay of the renal 1α -hydroxylase enzyme, and, consequently,

concentrations of $1,25(\text{OH})_2\text{D}$ may be inappropriately low until about the third week of extrauterine life. This, coupled with the limited absorptive capacity of the pre-term intestinal tract, contributes to the mineral substrate deficiency.

The biochemical changes in plasma that may identify those infants at risk of developing clinically significant osteopenia of prematurity are shown in Table 25.4. Plasma alkaline phosphatase activity fluctuates considerably during the neonatal period, but rises rapidly to values above six times the upper adult reference limit in the majority of infants at risk. High plasma enzyme activities are inversely related to growth velocity, a fall in which is an early clinical manifestation of the disorder. Plasma phosphate concentrations are invariably low-normal or low and urinary phosphate excretion is significantly reduced. Plasma calcium concentrations are a poor indicator of impending rickets and may be low, normal or even raised, depending upon the aetiology and stage of the disorder. Urinary calcium excretion is variable and related to the plasma phosphate concentration: the lower the plasma phosphate, the greater the urinary calcium excretion.

The treatment for this common form of osteopenia of prematurity is to increase calcium and phosphorus intake in order to maintain a normal plasma calcium concentration, but, more importantly, to keep the plasma phosphate close to about 1.50 mmol/L. Vitamin D supplementation, in an active form such as 1- α -calcitriol, is often administered. Following successful treatment, plasma alkaline phosphatase activity may continue to rise for about two weeks before falling to expected pre-term values.

Rickets during childhood

Rickets occurs in growing children as a result of defective mineralization of growth plates: it is similar to osteomalacia in adults. The major cause of rickets during childhood is nutritional vitamin D deficiency. Other causes include malabsorption and anticonvulsant therapy. Hypocalcaemia with secondary hyperparathyroidism is invariably present and children may present with tetany and convulsions. Older infants may present with a characteristic waddling gait.

Inherited forms of rickets must be considered in the differential diagnosis. They can be divided into two main groups: those associated with hypophosphataemia, caused by an isolated or generalized renal tubular defect, and those in which the main defect is in vitamin D metabolism (Box 25.4). X-linked dominant hypophosphataemic rickets is the most common form and occurs in about one in 20 000 live births. Hypophosphataemia, with isolated

BOX 25.4

Classification of inherited causes of osteopenia/rickets

Isolated renal tubular defect

- X-linked hypophosphataemic rickets
- Autosomal dominant/recessive hypophosphataemic rickets

Generalized renal tubular defect

- Renal tubular acidosis type 2
 - Cystinosis
 - Classic galactosaemia
 - Hereditary fructose intolerance
 - Tyrosinaemia type I
 - Wilson disease
- Renal tubular acidosis type 1

Defective vitamin D metabolism

- 1α -Hydroxylase deficiency
- End-organ resistance

hyperphosphaturia, usually presents during the first year of life and is associated with growth retardation. Remission may occur when the epiphyseal plates fuse and growth ceases, but the condition can relapse in later life. Female patients are less severely affected than males, the difference being caused by the random inactivation of one of the paired X chromosomes (lyonization). In females, presentation may be as osteomalacia in adulthood. A low-normal or low plasma concentration of $1,25(\text{OH})_2\text{D}$ in the presence of hypophosphataemia is suggestive of an altered feedback control on 1α -hydroxylase and impaired vitamin D metabolism.

Although rare, cystinosis is the most common cause of rickets in infants with type 2 RTA or Fanconi syndrome. Rickets develops by the second year of life and is probably caused by a combination of hypophosphataemia, systemic acidosis and impaired renal hydroxylation of vitamin D. In contrast, infants with type 1 RTA develop hypercalcaemia and nephrocalcinosis and may present clinically with polyuria and polydipsia, renal stones or rickets, usually before the fifth year of life. The diagnosis of cystinosis is confirmed by the finding an increased concentration of white cell cystine.

Plasma alkaline phosphatase activity in infancy

Plasma alkaline phosphatase (ALP) fluctuates considerably in the pre-term infant, and activities up to six times the adult reference limit are considered appropriate for age.

TABLE 25.4 Biochemical changes in osteopenia of prematurity and reference intervals compared with adults

Analyte (in plasma)	Units	Reference interval (Adults)	Reference interval (Pre-term infants, 3–10 weeks)	Change in osteopenia of prematurity
Alkaline phosphatase	U/L	30–130	<600	Raised
Phosphate	mmol/L	0.80–1.50	1.50–2.55	Low-normal or low
Calcium	mmol/L	2.20–2.60	2.15–2.65	Low, normal or high
Albumin	g/L	35–50	25–40	Variable

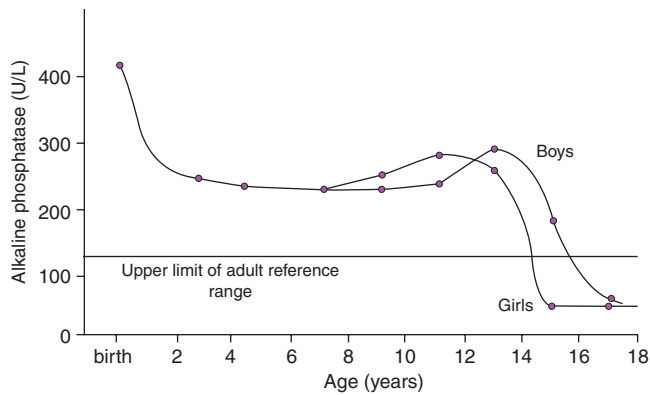


FIGURE 25.3 ■ Age- and sex-related changes in plasma alkaline phosphatase activity during childhood and adolescence.

Plasma ALP consists primarily of the bone isoform in infants, although fetal intestinal ALP, the gene for which is normally suppressed after the 30th week of gestation, may be present in the plasma of pre-term neonates receiving an oral fat intake. The hepatic isoform of ALP is rarely detected before one year of age, even in the presence of hepatic disease. Placental ALP is not present in the plasma of the newborn. Age and sex variation in plasma ALP during childhood and adolescence is shown in [Figure 25.3](#).

Transient hyperphosphatasaemia of infancy is a rare disorder of unknown aetiology that occurs in infants and children up to the age of about five years, although it may occasionally present in older children and adults. It is characterized by a rapid rise in plasma ALP activity to values >20 times the upper adult reference limit and a subsequent return to normal, usually within three months. There is no associated bone or liver disease and the disorder has no proven long-term sequelae. Its early recognition is important as it prevents unnecessary investigations (see Chapter 13).

DISORDERS OF LIVER FUNCTION

Bilirubin metabolism

Bilirubin is formed in the reticuloendothelial system of the liver, spleen and bone marrow from the breakdown of haemoglobin. Unconjugated bilirubin is transported in plasma bound to albumin, which has two bilirubin binding sites, one high and one low affinity, with a maximum binding capacity of about 20 μmol bilirubin/g albumin.

Bilirubin crosses the sinusoidal membranes of hepatocytes by non-energy-dependent passive diffusion. It becomes bound to Y protein (ligandin) and is transported to the smooth endoplasmic reticulum where it is conjugated, by the enzyme uridine diphosphate (UDP)-glucuronosyl transferase, to form a water-soluble mono- and then diglucuronide. The monoglucuronide is the predominant conjugate of bilirubin in the newborn, but only accounts for about 15% of total bilirubin conjugates in adults. The ability of the neonatal liver to conjugate bilirubin is limited (see below), but by two years of age, the conjugating capacity greatly exceeds the rate of delivery of bilirubin to the liver and only 1% of the enzyme activity is required to clear bilirubin at the normal rate of production.

Bilirubin conjugates are secreted across the canalicular membrane into the bile by a multispecific anion transporter. The rate of secretion of bilirubin into the biliary canaliculi is dependent on the active secretion of bile salts and the rate of bile flow. Up to 25% of the conjugated bilirubin may be hydrolysed within the intestinal lumen, either non-enzymatically, under alkaline conditions, or by the mucosal enzyme β -glucuronidase, to unconjugated bilirubin, which is reabsorbed and returned to the liver via the enterohepatic circulation. Conjugated bilirubin is converted to bile pigments (urobilinoids), primarily by intestinal bacteria in the distal ileum and colon.

Unconjugated hyperbilirubinaemia: physiological jaundice

Bilirubin bound to albumin in the blood may be displaced by drugs, such as antibiotics, and by hormones and free fatty acids. Free bilirubin is lipid soluble and can cross the blood–brain barrier, where its toxic effects result in irreversible brain damage (kernicterus). This is thought to be caused by uncoupling of oxidative phosphorylation. Significant free bilirubin occurs when the plasma total bilirubin concentrations exceed 340 $\mu\text{mol/L}$ in term infants, with a conjugated bilirubin <25 $\mu\text{mol/L}$. The absence of a clear relationship between circulating bilirubin concentrations and the development of kernicterus in pre-term infants has led to the ‘action level’ for total bilirubin being reduced to as low as 170 $\mu\text{mol/L}$. Putative additional risk factors for the development of kernicterus include acidosis, sepsis and hypothermia. National guidelines have been adopted for the detection and management of neonatal hyperbilirubinaemia to prevent the potential devastating effects. However, given that the cause may be a sudden and unpredictable haemolytic crisis, kernicterus has yet to be completely eliminated.

Hyperbilirubinaemia is usually treated with phototherapy, which converts the bilirubin to water soluble isomers but, if severe, may require treatment by exchange transfusion. This latter procedure may be complicated by a fall in plasma ionized calcium concentration and there is an associated risk of cardiac arrest.

Transient unconjugated hyperbilirubinaemia, presenting with jaundice on the second or third day of extra-uterine life and persisting for about two weeks, is known as ‘physiological jaundice of the newborn’. In premature infants, such jaundice may last for up to four weeks. During the first week of life, plasma bilirubin concentrations rise to >220 $\mu\text{mol/L}$ in 6% of healthy full-term neonates and a greater proportion of pre-term neonates. The cause is controversial and almost every phase of bilirubin metabolism and transport has been implicated. Bilirubin synthesis is increased during the first three weeks of life, particularly in premature infants, because of the reduced erythrocyte half-life. The high bilirubin production (100–140 $\mu\text{mol/kg}$ body weight/24h), almost three times the normal adult production, exceeds the capacity of the neonatal liver to clear unconjugated bilirubin from the plasma, owing to reduced membrane uptake and ligandin production, the low activity of UDP-glucuronosyl transferase and inefficient bile acid metabolism and transport.

After the third week of life, the increased bilirubin load is derived principally from the enhanced reabsorption of unconjugated bilirubin from the gut. The reduced uptake and conjugation of the increased load by the liver of the premature infant contributes to the prolonged 'physiological' jaundice in this age group. It has been suggested that bilirubin, through its oxidation to biliverdin, provides a scavenging mechanism for free radicals produced during disease and that physiological jaundice may have a beneficial role in the neonate.

Unconjugated hyperbilirubinaemia: pathological causes

Unconjugated hyperbilirubinaemia that occurs within 24h of birth, which rises rapidly ($>85 \mu\text{mol/L}/24\text{h}$) or which is persistent, should prompt investigation for the secondary conditions shown in **Box 25.5**. Haemolytic disease of the newborn is now uncommon following the widespread use of prophylactic rhesus immune globulin: when it occurs, jaundice develops within the first six hours of life. Congenital spherocytosis and ABO haemolytic disease are usually associated with less severe jaundice. The latter condition resolves as the circulating erythrocytes coated with anti-A and anti-B antibodies are broken down.

Erythrocyte glucose 6-phosphate dehydrogenase deficiency is a common inherited cause of haemolytic disease and should be ruled out in neonates with persistent hyperbilirubinaemia. Haemolysis in this condition may be precipitated by exposure to chemicals, drugs and infection and may be more severe in the Mediterranean and Oriental type than in that seen in African-Americans.

Congenital infections such as syphilis, rubella and toxoplasmosis may cause increased erythrocyte turnover and haemolysis, which can contribute to unconjugated hyperbilirubinaemia.

Inherited disorders of bilirubin metabolism presenting in childhood

Inherited glucuronosyl transferase deficiency (Crigler-Najjar syndrome) is a rare disorder of impaired bilirubin conjugation that presents with progressive unconjugated hyperbilirubinaemia from birth. In type I, there is complete absence of hepatic glucuronosyl transferase activity. The disorder is inherited as an autosomal recessive trait. With modern management of phototherapy (10–12 h/day) and exchange transfusions, kernicterus is now less common, but affected individuals still go on to develop

BOX 25.5 Pathological causes of unconjugated hyperbilirubinaemia in the neonate

- Rhesus haemolytic disease of the newborn
- ABO haemolytic disease
- Congenital spherocytosis
- Glucose 6-phosphate dehydrogenase deficiency
- Prenatal infection
- Hypothyroidism
- Inherited disorders of bilirubin metabolism

severe neurological complications during adolescence. Many individuals are treated by liver transplantation by the age of five years. Heterozygotes have normal plasma bilirubin concentrations but hepatic enzyme activity is reduced by about 50%. Some heterozygotes may present in the neonatal period with unconjugated hyperbilirubinaemia.

In type II, glucuronosyl transferase activity is detectable but markedly reduced. Unconjugated hyperbilirubinaemia presents at birth and persists into adult life. Milder forms of the disorder are indistinguishable from Gilbert syndrome, a benign autosomal dominant disorder characterized by mild fluctuating unconjugated hyperbilirubinaemia. Type II glucuronosyltransferase deficiency is inherited as an autosomal dominant trait with variable penetrance. As hepatic enzyme activity normally exceeds the rate of bilirubin delivery, small changes in enzyme activity can significantly alter the expression of the disorder. Treatment with phenobarbital, which induces enzyme activity, results in a rapid decline in plasma bilirubin concentration, maximal by the sixth day, in infants with the type II disorder but not with type I.

Conjugated hyperbilirubinaemia and hepatocellular disease

Conjugated hyperbilirubinaemia, defined in infants as a conjugated plasma bilirubin concentration of $>25 \mu\text{mol/L}$, is always pathological. As in adults, the cause may be extra- or intrahepatic cholestasis or hepatocellular necrosis (**Box 25.6**). Routine liver function tests are often poor discriminators of the aetiology during the neonatal period, but the presence of pale stools is suggestive of cholestasis.

The biliary atresias comprise a group of disorders in which all grades of bile duct obstruction may be present, from a marked reduction in bile duct numbers to complete absence. Conjugated hyperbilirubinaemia develops during the first week of life and continues to increase, causing a fluctuating jaundice and failure

BOX 25.6 Some causes of neonatal/childhood liver disease

Neonatal

- Cholestasis
 - Biliary atresia
 - Choledochal cyst
 - α_1 -Antitrypsin deficiency
 - Defects of bile acid synthesis
 - Parenteral nutrition
- Hepatocellular disease
 - Infection
 - Neonatal haemochromatosis
 - Inherited metabolic diseases (Zellweger syndrome, tyrosinaemia type I, classic galactosaemia, fructose intolerance, Niemann–Pick type C)

Childhood

- Viral hepatitis
- α_1 -Antitrypsin deficiency
- Wilson disease
- Cystic fibrosis

to thrive. Early diagnosis, by abdominal ultrasound and isotope studies, is important, as extrahepatic atresia can be successfully treated surgically with the Kasai procedure (hepatic portoenterostomy). Liver transplantation is undertaken in some cases.

α_1 -Antitrypsin deficiency is probably the most common inherited metabolic disorder to cause neonatal conjugated hyperbilirubinaemia sufficient to mimic severe biliary atresia. Liver disease occurs in those patients with the PiZZ phenotype, of whom 10–20% present with neonatal cholestasis. The diagnosis must be confirmed by determining the phenotype/genotype.

Defects in the synthesis of bile acids (cholic acid and chenodeoxycholic acid) from cholesterol are rare inherited causes of cholestatic jaundice. The jaundice is usually associated with steatorrhoea and malabsorption of the fat-soluble vitamins. Defining defects of bile acid synthesis depends upon the identification of unusual profiles of bile acids or alcohols in urine.

Prolonged parenteral nutrition may cause cholestatic liver disease with a progressive rise in plasma alkaline phosphatase activity and a later rise in the plasma aminotransferases. The aetiology is unclear but it may be caused by an ascending cholangitis, secondary to intestinal stasis. The condition usually resolves when parenteral feeding is stopped and enteral feeding is started or resumed but, in some children, the abnormalities persist and chronic liver disease may supervene. Some children have been successfully treated with combined liver and small bowel transplantation.

Hepatocellular disease in neonates is associated with raised plasma aminotransferase activities, reduced plasma albumin and prolonged prothrombin time. Clinically, there may be oedema and ascites. When liver damage has occurred in utero, for instance in congenital infections and some inherited metabolic diseases, infants may present soon after birth with the neurological effects of intracranial bleeding before liver disease becomes apparent clinically.

Postnatally acquired infection may be superimposed on an inherited metabolic disease, such as tyrosinaemia type I or classic galactosaemia.

Neonatal haemochromatosis is a rare inherited disorder characterized by jaundice, hepatic failure and hypoglycaemia in the newborn period. Although the distribution of iron overload in the liver and in extrahepatic tissues, such as the heart and pancreas, is similar to that of hereditary haemochromatosis, the two conditions are distinct. The aetiology of neonatal haemochromatosis is unclear; it is probably inherited as an autosomal recessive trait. The transferrin saturation (80–90%) and serum ferritin concentration are both markedly raised.

In many infants, it is impossible to differentiate cholestasis and hepatocellular disease. Several inherited metabolic diseases are associated with neonatal hepatocellular disease. Sometimes there are clinical pointers to a diagnosis, such as the craniofacial dysmorphism of the peroxisomal disorder Zellweger syndrome (p. 462). In the absence of any specific clinical signs, the following investigations should be performed:

- blood: gases, $[H^+]/pH$
- plasma: lactate, glucose, ammonia, amino acids, α_1 -antitrypsin, prothrombin time (international normalized ratio).

Given that the liver has significant metabolic turnover, it is not surprising that inherited metabolic diseases, such as fatty acid oxidation defects, glycogen storage disease type I and urea cycle disorders, are associated with early liver disease. Affected infants are acutely ill and the diagnosis may be indicated by hypoglycaemia or hyperammonaemia, in addition to deranged liver function tests.

Disorders in which the major initial biochemical findings relate to the liver include the acute form of tyrosinaemia type I (fumarylacetoacetate hydrolase deficiency). Signs and symptoms of the acute neonatal onset form include acute liver failure with hepatomegaly, a tubulopathy, hypoglycaemia, failure to thrive, vomiting, diarrhoea and a cabbage-like odour; some infants may present with a neurological crisis. Plasma tyrosine and methionine concentrations are markedly raised. Urine organic acid analysis shows raised concentrations of the tyrosine metabolites, 4-hydroxyphenyllactic and 4-hydroxyphenylpyruvic acids, and succinylacetone. Of these biochemical findings, only the latter, succinylacetone, is specific to the defect, and allows differentiation from transient tyrosinaemia of infancy. Tyrosinaemia type I can now be successfully treated using 2-(2-nitro-4-trifluoromethylbenzoyl)-1,3-cyclohexanedione (NTBC), which inhibits the enzyme 4-hydroxyphenylpyruvate dioxygenase and thus the formation of fumarylacetoacetate and succinylacetone. Succinylacetone inhibits the enzyme porphobilinogen synthase; consequently, older individuals with tyrosinaemia type I may present with features resembling those of acute intermittent porphyria. Hepatocellular carcinoma is a significant complication, and patients should be monitored regularly by measuring plasma α -fetoprotein.

There are three separate inherited disorders of galactose metabolism. The most common and severe of these (classic galactosaemia) is galactose 1-phosphate uridylyltransferase (GALT) deficiency. The acute clinical presentation of galactosaemia includes vomiting and diarrhoea, hepatic dysfunction, with a pronounced coagulopathy, an *Escherichia coli* septicaemia due to impaired leukocyte bactericidal activity, and cataracts. Galactose 1-phosphate uridylyltransferase activity should be measured in red blood cells.

Dietary galactose restriction can reverse these acute problems, but the development of long-term complications, such as developmental delay, speech abnormalities and premature ovarian failure, may not be affected by the early institution of therapy. The poor treatment outcome in classic galactosaemia is due either to endogenous production of galactose 1-phosphate through the action of epimerase or to the low intracellular availability of UDP-galactose, a necessary component of some glycolipids and glycoproteins.

Liver disease in older children

The most common causes of liver disease in childhood are viral, autoimmune and drug-induced hepatitis. However, some metabolic diseases, such as α_1 -antitrypsin deficiency, tyrosinaemia, Wilson disease or cystic fibrosis should be excluded. A Fanconi-like picture with osteopenia and rickets may indicate the presence of renal tubular damage secondary to tyrosinaemia type I, Wilson disease

or hereditary fructose intolerance. Tyrosinaemia type I may also present with hypertension and a porphyria-like crisis, owing to succinylacetone-induced inhibition of 5-aminolaevulinic acid dehydratase (ALAD), the rate-limiting enzyme in the porphyrin metabolic pathway.

Wilson disease

This recessively inherited disorder is caused by the impaired hepatic incorporation of copper into caeruloplasmin and by reduced biliary excretion of copper. Copper accumulates in the liver and later in the brain and kidneys. The disorder may present with acute liver failure, cirrhosis or chronic hepatitis after four years of age but asymptomatic hepatocellular damage has been found earlier in life. A more detailed account of Wilson disease is given in Chapter 14.

Reye syndrome or Reye-like illness

Hepatitis associated with acute encephalopathy and fatty infiltration of tissues is known as Reye syndrome. The onset of the disorder may be precipitated by a variety of conditions, including viral illness, especially varicella or influenza A or B, drugs, such as salicylates or sodium valproate, and toxins, including insecticides and aflatoxins.

The condition usually presents in children aged 3–12 years. Following a viral-like illness, the child develops persistent vomiting and a progressive encephalopathy, with irritability, confusion and, in its most severe form, coma. Jaundice is rarely present. Biochemical features include hypoglycaemia, raised plasma ammonia concentration and aminotransferase activity and a prolonged prothrombin time.

Several inherited metabolic disorders present clinically and biochemically in a manner similar to Reye syndrome. These include disorders of medium and long chain fatty acid oxidation and some organic acidaemias. Reye syndrome should not be diagnosed in a child <3 years of age until a genetic defect has been excluded. Given the possible link between Reye syndrome and salicylates it is recommended that they are not given to children under

16 years. This has led to a significant drop in the number of cases recorded.

Further reading

- Blau N, Duran M, Blaskovics ME et al. *Physician's guide to the laboratory diagnosis of metabolic diseases*. 2nd ed. Berlin: Springer-Verlag; 2003.
- Green A, Morgan I, Gray J. *Neonatology and laboratory medicine*. 2nd ed. London: ACB Venture Publications; 2003.
- A succinct, inexpensive book for the biochemist covering primarily the laboratory diagnosis and management of disorders presenting in the neonate. The book also includes a number of illustrative cases and specific investigation protocols.*
- Mérelle ME, Nagelkerke AF, Lees CM et al. *Newborn screening for cystic fibrosis*. *Cochrane Database Syst Rev* 2001;3, CD001402.
- Milford-Ward A, Sheldon J, Rowbottom A et al. editors. *PRU handbook of clinical immunochemistry*. 9th ed. Sheffield: PRU Press; 2007.
- Rennie JM, editor. *Rennie & Robertson's textbook of neonatology*. 5th ed Edinburgh: Churchill Livingstone; 2012.
- An extensive textbook covering all aspects of clinical neonatology including laboratory investigations; this is an important reference book for clinical biochemists working closely with neonatal units.*
- Scriver CR, Beaudet AL, Sly WS et al. editors. *The metabolic and molecular bases of inherited disease*. 8th ed New York: McGraw-Hill; 2001.
- The essential standard reference book for those with a special interest in inherited metabolic disorders.*
- Zschocke J, Hoffmann GF. *Vademecum metabolicum. Manual of metabolic paediatrics*. 2nd ed. London: Schattauer; 2004.

Internet resources [All Accessed October 2013]

- Congenital disorders of glycosylation, <http://www.euroglycanet.org>.
- Mitochondrial disorders, <http://www.neuro.wustl.edu/neuromuscular/mitosyn.html>.
- Neuromuscular Disease Center, Washington University, St Louis, MO, USA. <http://neuromuscular.wustl.edu/>.
- Online Mendelian Inheritance in Man. <http://www.ncbi.nlm.nih.gov/omim>.
- Catalogue of human genes and genetic disorders maintained by the National Centre for Biotechnology Information, USA.*
- UK Metabolic Biochemistry Laboratory Network. <http://methbio.net>.
- Assay directory for specialist metabolites and enzymes for inherited metabolic disorders. There is also an active training and education initiative and best practice guidelines aimed at helping local non-specialist laboratories and clinical teams.*
- UK Newborn Screening Programme Centre. <http://newbornscreening-bloodspot.org.uk>.
- Website includes details of the UK screening programme and UK standards.*

Introduction to haematology and transfusion science

David Ah-Moye • Ceinwen Davies • Joanne Goody • Peter Hayward • Rebecca Frewin

CHAPTER OUTLINE

INTRODUCTION 497

GENERAL HAEMATOLOGY 497

- Analysis of the full blood count 497
- Reticulocyte count 499
- Erythrocyte sedimentation rate and plasma viscosity 499
- Flow cytometry 499
- Haematinic studies 500
- Haemoglobinopathy screening 500
- Tests for infectious mononucleosis 500

MORPHOLOGY 501

- Blood film examination 501
- Normal red cell morphology 501
- Morphology of the anaemias 501
- Normal white cell morphology 503
- Abnormal white cell morphology 504
- Haematological malignancies 504

HAEMOSTASIS 507

- The coagulation cascade 507
- Laboratory tests of coagulation 507
- Interpretation of coagulation tests 509

BLOOD TRANSFUSION 510

- Introduction 510
- Blood group antigens 510
- Laboratory transfusion tests 511
- Investigation of suspected transfusion reaction 512
- Haemolytic disease of the newborn 512
- Blood products 513
- Risks of transfusion 513
- Regulations 514

CONCLUSION 514

INTRODUCTION

Haematology is the study of blood, its formation, composition, functions and diseases. Blood cell formation (haemopoiesis) is described in Chapter 27. Haematology laboratories perform a wide range of tests of the composition and function of the blood that assist the clinician in the diagnosis and management of disease. They are typically divided into three principal sections: general haematology, haemostasis and blood transfusion. This chapter provides an overview of the different diagnostic techniques offered in each.

GENERAL HAEMATOLOGY

The general haematology laboratory performs numerical, morphological and biochemical analysis of blood samples (now often as part of a combined blood sciences laboratory). Blood samples must be anticoagulated, for example with ethylenediamine tetra-acetic acid (EDTA) for determination of full blood count and reticulocyte count, blood film examination and plasma viscosity and

tri-sodium citrate for coagulation tests and measurement of erythrocyte sedimentation rate.

Analysis of the full blood count

The full blood count (FBC) is the most requested test profile in haematology. It comprises:

- haemoglobin (Hb) concentration
- red blood cell (RBC) count
- red cell indices (volume and haemoglobin content)
- white blood cell (WBC) count
- white blood cell differential count (neutrophils, lymphocytes, monocytes, eosinophils and basophils)
- platelet count.

The FBC is a valuable first-line screening test from which further investigations may be generated (see [Table 26.1](#)). The clinician uses information from the FBC in conjunction with the clinical history, physical examination and other investigations to help formulate and distinguish between differential diagnoses.

Full blood counts are usually performed on fully automated analysers, which provide accurate quantitation of cell parameters and can generate alerts, commonly

TABLE 26.1 Haematology adult typical reference ranges (based on those used in the authors' laboratory)

Parameter	Units	Range (females)	Range (males)
Hb	g/L	115–165	130–180
RBC	10 ¹² /L	3.80–5.80	4.50–6.50
HCT	L/L	0.37–0.47	0.40–0.54
MCV	fL	80.0–100.0	80.0–100.0
MCH	pg	27.0–32.0	27.0–32.0
MCHC	g/L	285–330	285–330
WBC	10 ⁹ /L	3.6–11.0	3.6–11.0
PLT	10 ⁹ /L	140–400	140–400
Neutrophils	10 ⁹ /L	1.8–7.5	1.8–7.5
Lymphocytes	10 ⁹ /L	1.0–4.0	1.0–4.0
Monocytes	10 ⁹ /L	0.2–0.8	0.2–0.8
Eosinophils	10 ⁹ /L	0.1–0.4	0.1–0.4
Basophils	10 ⁹ /L	0.01–0.1	0.01–0.1
ESR (age <50)	mm/h	3–9	1–7
ESR (age ≥50)	mm/h	5–15	2–10
Plasma viscosity	mPa/s	1.50–1.72	1.50–1.72
Reticulocytes	10 ⁹ /L	76–130	76–130
Serum B ₁₂	ng/L	180–1000	180–1000
Serum folate	µg/L	>4.0	>4.0
Serum ferritin	µg/L	10–300	25–350

All values are for whole blood unless stated. Hb, haemoglobin; RBC, red blood cells; HCT, haematocrit; MCV, mean cell volume; MCH, mean cell haemoglobin; MCHC, mean cell haemoglobin concentration; WBC, white blood cells; PLT, platelets; ESR, erythrocyte sedimentation rate.

known as 'flags', to the presence of unusual characteristics that may warrant review of a blood film.

Haemoglobin

The main function of haemoglobin, which is present solely within the red cells, is to transport oxygen from the lungs to the tissues (see Chapter 5). Haemoglobin is measured by spectrometry. Diluted whole blood is mixed with a reagent (e.g. Stromatolyser™) that results in cell lysis and conjugates the haemoglobin with a compound such as potassium cyanide or sodium lauryl sulphate prior to measurement of its absorbance. The presence of high concentrations of substances that increase absorbance, e.g. triglyceride-rich lipoproteins, may cause a falsely raised haemoglobin result with some methods.

Cell counting

Blood cells are counted and sized either by electrical impedance or laser light scattering methods. The impedance method utilizes the fact that blood cells are very poor conductors of electricity. When a cell passes through a small aperture across which an electric current is applied, there is a measurable rise in the electrical impedance, with the pulse height being proportional to the volume of the cell. Each cell generates a separate impulse. Light scatter utilizes the fact that cells scatter light when passed through a laser beam: the number of cells can be counted and information about the cells can be obtained according to the pattern of scatter. White cell counting is performed after lysis of the red cells.

Clumping of cells results in underestimation of cell numbers. This may occur, for example, as a result of autoagglutination of red cells in cold agglutinin disease or in vitro clumping of platelets because of inadequate mixing with EDTA anticoagulant.

Red cell indices

The red cell indices are a combination of measured and derived red cell parameters. The red cell indices usually reported are:

- mean cell volume (MCV)
- haematocrit (HCT) or packed cell volume (PCV), the percentage volume of red cells in the blood
- mean cell haemoglobin (MCH), the average mass of haemoglobin per red cell
- mean cell haemoglobin concentration (MCHC), the average concentration of haemoglobin per red cell.

Some analysers measure the MCV and calculate the HCT or PCV; others measure the HCT and calculate the MCV.

Mean cell volume is useful when assessing anaemia. For example, it is reduced in iron deficiency anaemia and increased in megaloblastic anaemia.

Mean cell haemoglobin is reduced in iron deficiency anaemia and thalassaemias; further tests such as ferritin and HbA₂ measurement may be performed to differentiate between these two conditions.

Mean cell haemoglobin concentration is reduced in iron deficiency but is probably a more useful tool within the laboratory than clinically. It is very stable in health, so it can be used as an internal quality control; a significant change in the daily mean MCHC indicates a drift in analyser calibration or a fault in its function. The MCHC is also useful for detecting anomalous results: a falsely raised Hb concentration, for example as a consequence of lipaemia, or a falsely reduced RBC count, for example owing to red cell autoagglutination, will result in a falsely raised MCHC.

The formulae for the derived parameters are as follows:

$$\text{MCV (fL)} = \frac{\text{HCT (L/L)} \times 1000}{\text{RBC (10}^{12}\text{/L)}}$$

$$\text{MCH (pg)} = \frac{\text{Hb (g/L)}}{\text{RBC (10}^{12}\text{/L)}}$$

$$\text{MCHC (g/L)} = \frac{\text{Hb (g/L)}}{\text{HCT (L/L)}}$$

White cell differential

The white blood cells are involved in the cellular immunological response to infection and other causes of inflammation. The five types of white blood cell differ in volume, conductivity, light scattering properties, uptake of fluorescent dyes and cytochemical stains, and resistance to lytic agents. These properties, used in various

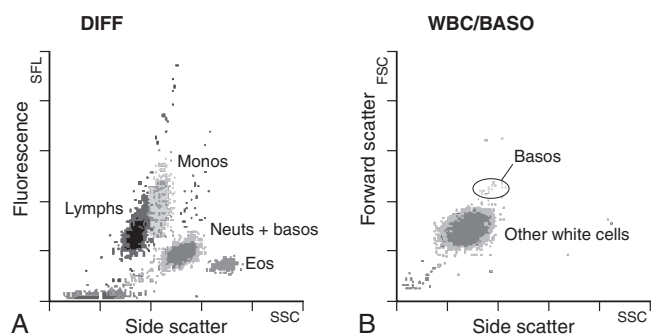


FIGURE 26.1 ■ Five-part WBC differential from Sysmex XE2100™ analyser. Note that neutrophils and basophils cannot be separated in the side fluorescence (SFL) vs side scatter (SSC) population scattergram (A). The basophil count is obtained by treating the white cells with an acidic reagent that shrinks all white cells, leaving bare nuclei, except for basophils, which remain intact (B). The neutrophil count is then derived by subtracting the basophil count from the combined neutrophil and basophil count. Lymphs, lymphocytes; monos, monocytes; neuts, neutrophils; eos, eosinophils; basos, basophils.

combinations, enable their identification. **Figure 26.1** shows typical white cell scattergrams for a normal blood sample.

Platelet count

The main function of platelets is to form a plug at sites of damage to vascular endothelium. A low platelet count (thrombocytopenia) may result from decreased production, increased consumption or abnormal pooling in the spleen. Decreased production is a consequence of either general marrow failure or selective depression, for example by viral infection, drugs and chemicals. Platelet consumption is a feature of certain types of autoimmune diseases and occurs in disseminated intravascular coagulation (which is often a result of septicaemia) and thrombotic thrombocytopenic purpura. A high platelet count (thrombocytosis) may be seen as part of a reactive process such as infection or in myeloproliferative neoplasms. Automated haematology analysers routinely measure the platelet count by impedance or laser light scattering; the reference method uses flow cytometry after labelling the platelets with CD41 or CD61 antibodies.

Reticulocyte count

Reticulocytes are immature red cells, which are released from the bone marrow into the circulation one to two days before maturation. An increase in the reticulocyte count is indicative of an increase in red cell production (erythropoiesis) to meet physiological demand, for example in haemolytic anaemia.

Reticulocytes contain ribosomal RNA, which can be detected by two main methods. When unfixed red cells are incubated with dyes such as new methylene blue, the ribosomal RNA precipitates out and appears as a blue reticular network within the cells, which can be visualised using a light microscope, allowing the reticulocytes to be counted. Most automated FBC analysers now perform reticulocyte counts by flow cytometry, using a fluorescent dye that binds to the ribosomal RNA; the number of

reticulocytes can be expressed as a percentage of total red cells and as an absolute count.

Erythrocyte sedimentation rate and plasma viscosity

Erythrocyte sedimentation rate (ESR) expressed in mm/h, is a measure of the rate at which red blood cells settle when an anticoagulated blood sample is left to stand in a column.

Plasma viscosity is usually measured by an automated viscometer. The plasma is drawn through a capillary tube under constant pressure and temperature; the rate of flow is measured and expressed as viscosity of plasma relative to distilled water.

Measurement of ESR or plasma viscosity may be useful when assessing the acute phase response. Both are affected by the presence of large proteins such as fibrinogen, α_2 -macroglobulin and immunoglobulins, and are therefore usually raised during an acute phase response. They are non-specific screening tools for inflammation, and are useful in monitoring of diseases such as rheumatoid arthritis and response to treatment for example with non-steroidal anti-inflammatory drugs and anti-tumour necrosis factor.

Plasma viscosity has several advantages over ESR. For example, it is not affected by physiological and environmental factors. Measurement is rapid, reproducible and lends itself better to quality assurance procedures than measurement of ESR. Conversely, ESR is a cheap test, but is influenced by factors such as anaemia, sex, age and (in females) stage of the menstrual cycle and must be carried out within about six hours of venepuncture.

Flow cytometry

White cells express characteristic nuclear, cytoplasmic and surface antigens; this is referred to as the immunophenotype of the cell. Immunophenotyping has been made possible by the availability of a large number of antibodies, which target specific cellular antigens. For example, antibodies that are directed against the CD19 antigen will bind to B lymphocytes.

Immunofluorescence forms the basis of flow cytometric investigations. It uses antibodies conjugated to fluorescent dyes (fluorochromes) such as fluorescein isothiocyanate (FITC), phycoerythrin (PE) and phycoerythrin cyanin (PC). Whole blood or bone marrow is incubated with one or more antibodies, each conjugated to a different fluorochrome. The cells are then injected in single file into the path of a laser beam. As the cell passes through, the light is scattered in all directions. Forward scatter (FSC) is influenced by cell size whereas side scatter (SSC) is influenced by the cellular structure and contents. The different cell populations can be visualized by plotting SSC against FSC. As the cell-antibody complex passes through the laser beam it fluoresces. The different wavelengths of light emitted are filtered, enabling specific cell populations to be selected (gated) and the data can be displayed as dot plots or histograms. The dot plots show the intensity of fluorescence as well as the number of cells

that are positive for the particular antigen; the intensity of fluorescence is proportional to the strength of antigen expression. In the example shown in [Figure 26.2](#), whole blood was incubated with three antibody-dye conjugates, CD8-FITC, CD4-PE and CD3-PC5.

Flow cytometry allows rapid immunophenotyping of cells. This is essential in the diagnosis and classification of leukaemias and lymphomas, but is also a useful tool in the assessment of non-malignant conditions and for monitoring the CD4+ lymphocyte count during the treatment of human immunodeficiency virus (HIV) infection. Flow cytometry can also be used to confirm feto-maternal haemorrhage (see [p. 512](#)), by the detection of Rh D positive fetal cells and measurement of haemoglobin F in the fetal cells.

Haematinic studies

These comprise direct measurement of the concentration of ferritin (a measure of iron status), vitamin B₁₂ and folate. They are discussed further in [Chapter 27](#).

Haemoglobinopathy screening

The haemoglobinopathies and the techniques used for their diagnosis are discussed in [Chapter 29](#).

Tests for infectious mononucleosis

Infectious mononucleosis (IM), also known as glandular fever, is caused by the Epstein-Barr virus (EBV), which infects B lymphocytes. It is associated with a high titre of a heterophile antibody, so-called because it also reacts with red cells of other species, for example sheep, ox and horse. This antibody has been termed the Paul-Bunnell antibody, named after John Paul and Walls Bunnell who discovered them.

The Paul-Bunnell antibody is not adsorbed by guinea pig kidney, in contrast to other heterophile antibodies (Forssman antibodies). Differential adsorption tests utilize this principle; the patient's serum is incubated with ox red cells and separately with guinea pig kidney antigen, after which horse red cells (reagent) are added to the mixtures. Red cell agglutination only in the latter mixture indicates the presence of the Paul-Bunnell antibody.

An alternative test is the latex agglutination test. A reagent containing latex microspheres coated with purified Paul-Bunnell antigen is mixed with the patient's serum; agglutination indicates presence of the Paul-Bunnell antibody.

Definitive confirmation of EBV infection requires serological detection of a rise in the titre of specific

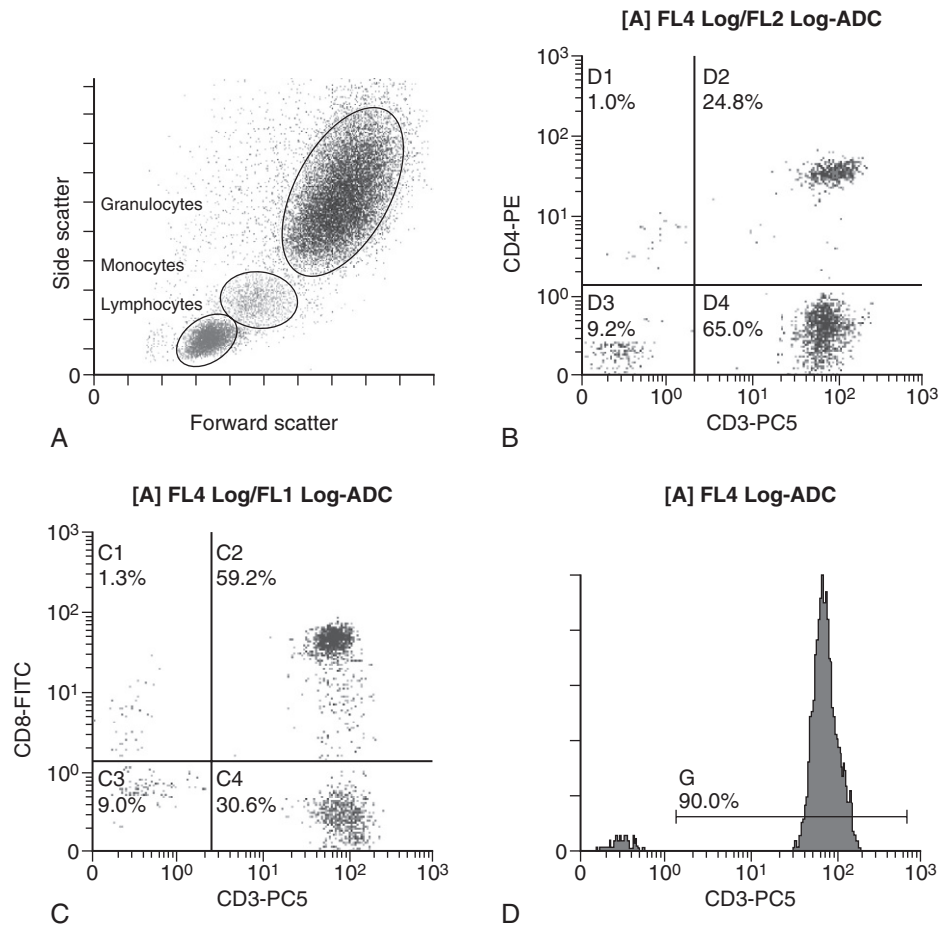


FIGURE 26.2 ■ Scatter plots, dot plots and histogram printouts from Beckman Coulter FC500 flow cytometer. (A) Side scatter vs forward scatter, showing white cell populations. (B) CD4 vs CD3 dual dot plots. (C) CD8 vs CD3 dual dot plots. (D) Single histogram for CD3 gated for the lymphocyte population. Note: Granulocytes is the collective name for neutrophils, eosinophils and basophils.

anti-EBV capsid antigen IgM, which occurs in the first few weeks following infection.

MORPHOLOGY

Blood film examination

Microscopic examination of a well stained blood film is an essential part of many haematological investigations. The film may be made either manually or by an automated slide maker. After being air dried, it is stained with Romanowsky dyes such as May–Grünwald Giemsa or modified Wright stain, which contain azure B and eosin. This enables the assessment of the size, shape, maturity and contents of blood cells by light microscopy. Blood film examination is useful in the investigation of the causes of anaemia and to diagnose white cell disorders such as leukaemia and parasitic infections such as malaria. Less commonly, a blood film may be requested for specific reasons such as to look for acanthocytes (red cells with irregular spicules), which are associated with abetalipoproteinaemia, liver disease and several rare degenerative neurological diseases.

Normal red cell morphology

In health, there is little variation in the shape and size of red blood cells. On a well spread and stained blood film, red cells appear circular in outline, with a well stained outer rim and a paler central area that occupies approximately a third of the cell (see Fig. 26.3). The normal red cell appears slightly smaller than the nucleus of a small lymphocyte. It is referred to as normocytic (normal size) and normochromic (normal staining characteristics).

Valuable information can be obtained from the shape of a red cell, as shown in Figure 26.4. In many haematological conditions there may be an increase in the proportion of non-specifically abnormal red cells, termed poikilocytosis. It is also good practice to check the patient's biochemistry results, to corroborate the morphological

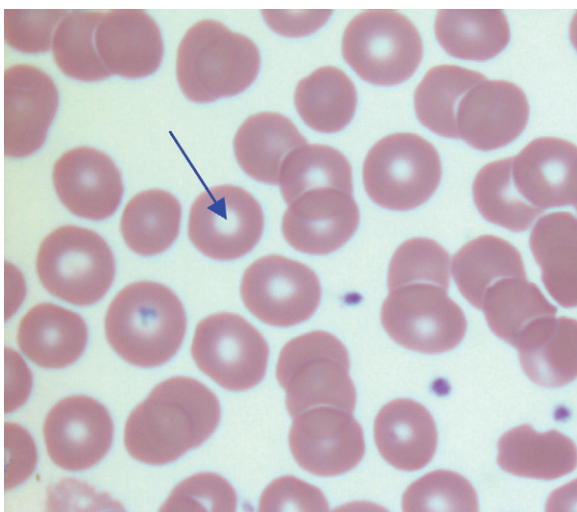


FIGURE 26.3 ■ Normal red cells. Shows area of central pallor (arrow) occupying one-third the size of the cell.

features; for example, if acanthocytes are present, the liver function tests should be checked, as acanthocytes are commonly present in liver disease.

Morphology of the anaemias

Anaemia is a condition in which the number of red blood cells, or their oxygen-carrying capacity, is insufficient to meet physiological needs. This is conventionally defined as a reduction in haemoglobin concentration below the reference range, i.e. <130 g/L in adult males and <115 g/L in adult females. The causes and assessment of anaemia are discussed in Chapter 27.

Iron deficiency anaemia

In iron deficiency anaemia, the red cells are smaller than normal (microcytosis). This is because the maturing red cells undergo an extra cellular division before the critical haemoglobin concentration required to arrest mitosis is achieved. The cells are also hypochromic, with a larger area of central pallor (see Fig. 26.5). Polychromatic cells (immature red cells that have a bluish hue on routine staining because of retained ribosomal ribonucleic acid) may be present, indicating increased erythropoiesis.

Megaloblastic anaemia

Megaloblastic anaemia is commonly caused by deficiency of vitamin B₁₂ or folate, both of which are essential for DNA synthesis, or the administration of drugs that interfere with DNA synthesis (e.g. methotrexate). Defective DNA synthesis results in the nucleus maturing at a slower rate than the cytoplasm, thereby producing a red cell that is larger than normal (macrocyte). Teardrop cells and red cell fragments may also be seen, as a consequence of ineffective erythropoiesis (see Fig. 26.6). A common finding is the presence of hypersegmented neutrophils (defined as presence of a nucleus with more than five lobes).

Autoimmune haemolytic anaemia

Haemolytic anaemia is a term used to describe anaemia caused by a shortened red cell lifespan. In autoimmune haemolytic anaemia, the red cells are coated with antibody (usually IgG) that is directed against the patient's own cells. When these cells circulate through the reticuloendothelial system, the surface of the red cells is eroded, leading to a change in their shape from a biconcave disc to a sphere (see Fig. 26.7).

Microangiopathic haemolytic anaemia

Microangiopathic haemolytic anaemia is a term that is used to describe the anaemia that results from physical damage to the red cells following the occlusion of arterioles and capillaries as a result of fibrin deposition or platelet aggregation. There are numerous causes, including infections (resulting, for example in disseminated intravascular coagulation (Fig. 26.8) or haemolytic uraemic syndrome), physical trauma (e.g. from mechanical heart valves) and autoimmune (e.g. thrombotic thrombocytopenic













Red cell abnormality	Possible causes	Red cell abnormality	Possible causes
 Target cell	Iron deficiency, liver disease, post-splenectomy, haemoglobinopathies, e.g. thalassaemia	 Elliptocyte	Hereditary elliptocytosis, some haemoglobinopathies, iron deficiency, myelofibrosis, megaloblastic anaemia
 Stomatocyte	Liver disease, alcoholism	 Ovalocyte	S.E. Asian ovalocytosis, macro ovalocytes – impaired DNA synthesis, e.g. B12 and folate deficiency
 Sickle cell	Diseases associated with the presence of haemoglobin S, e.g. sickle cell disease, haemoglobin SC disease	 Keratocyte (horned cell)	Microangiopathic haemolytic anaemia, e.g. disseminated intravascular coagulation, thrombotic thrombocytopenic purpura
 Tear drop cell	Myelofibrosis, metastatic marrow infiltration, megaloblastic anaemia	 Schistocyte (red cell fragment)	Microangiopathic and mechanical haemolytic anaemia
 Spherocyte; no area of central pallor	Immune haemolytic anaemia, post-splenectomy, hereditary spherocytosis, severe burns	 Echinocyte (crenated red cell); short, regular spicules	Multi-organ failure including hepatic and renal failure. Most commonly an artefact when a blood film is made from an old sample
 Irregularly contracted red cell; similar to spherocyte but not round	Haemoglobinopathies, oxidant damage especially in G6PD deficiency	 Acanthocyte; irregular spicules	Liver disease, abetalipoproteinaemia, inherited acanthocytosis

FIGURE 26.4 ■ Red blood cells. Commonly observed morphological changes in red cells and the disorders in which they arise.

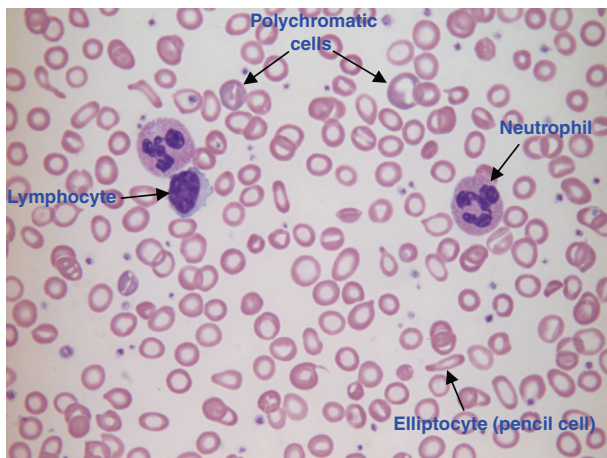


FIGURE 26.5 ■ Iron deficiency anaemia (IDA). Hb 58g/L, MCV 52.1 fL, MCH 12.7pg, PLT $608 \times 10^9/L$, ferritin $<5 \mu\text{g/L}$. Blood film shows microcytic hypochromic red cells, polychromatic red cells and pencil cells. The film also shows two neutrophils, a lymphocyte and an increase in platelet numbers.

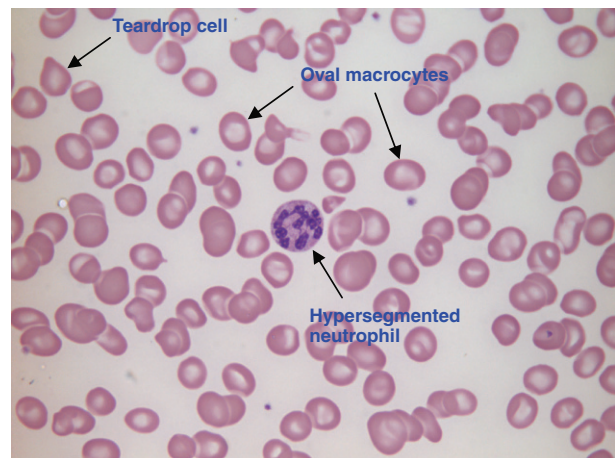


FIGURE 26.6 ■ Megaloblastic anaemia Hb 86g/L, MCV 121.9fL, MCH 43.9pg, PLT $110 \times 10^9/L$, Vit B₁₂ 62ng/L. Blood film shows macrocytic red cells, including oval macrocytes. The film also shows teardrop cells, a hypersegmented neutrophil and reduced platelet numbers.

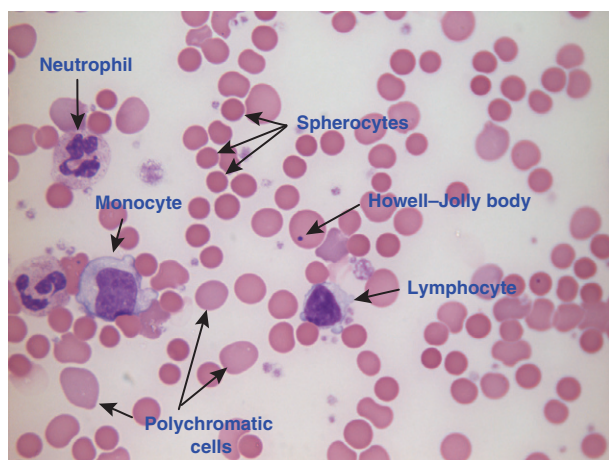


FIGURE 26.7 ■ Warm autoimmune haemolytic anaemia. Hb 104 g/L, reticulocytes $604 \times 10^9/L$, haptoglobin <0.2 g/L (reference range 0.7–3.2), direct antiglobulin test positive. Blood film shows numerous spherocytes, polychromatic cells and a red cell containing a Howell–Jolly body. Howell–Jolly bodies are typically found in patients who have undergone splenectomy. The film also shows two neutrophils, a monocyte and a lymphocyte.

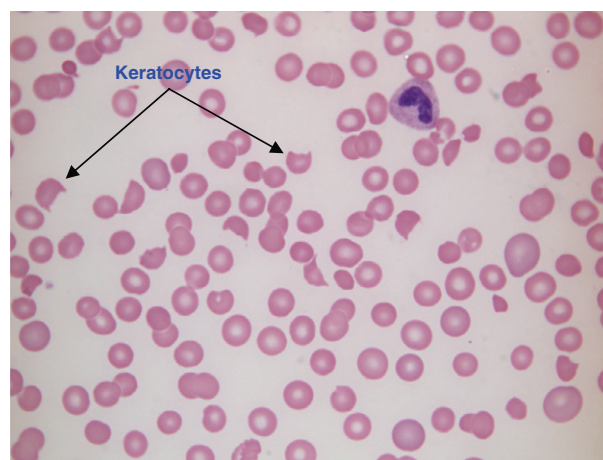


FIGURE 26.9 ■ Haemolytic uraemic syndrome. Hb 84 g/L, WBC $24.3 \times 10^9/L$, PLT $23 \times 10^9/L$. Blood film shows numerous red cell fragments, some with the characteristic horn shape (keratocyte) and reduced platelet numbers. The WCC is increased because of neutrophilia as a result of bacterial infection with *E. coli*.

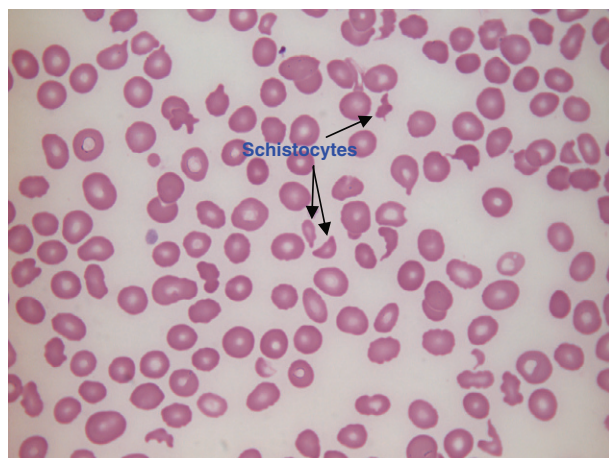


FIGURE 26.8 ■ Disseminated intravascular coagulation (DIC). Hb 89 g/L, WBC $3.0 \times 10^9/L$, PLT $77 \times 10^9/L$. Blood film shows numerous schistocytes and reduced platelet numbers.

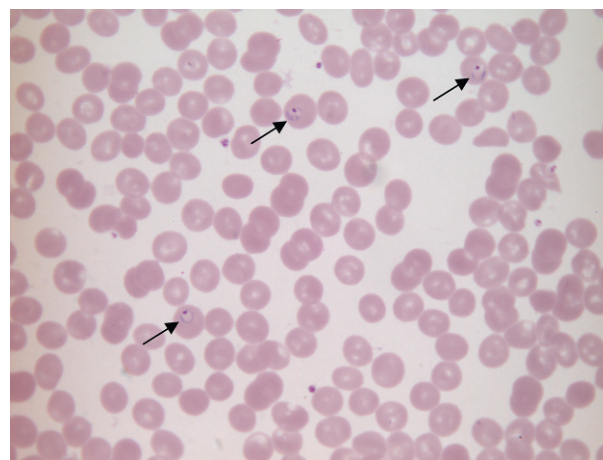


FIGURE 26.10 ■ *Plasmodium falciparum* malaria. Hb 149 g/L, WBC $6.1 \times 10^9/L$, PLT $111 \times 10^9/L$. Blood film stained by the dilute Giemsa method, showing *Plasmodium falciparum* trophozoites (ring forms) in three red cells (arrows).

purpura). The blood film contains numerous fragmented cells (called schistocytes), and immature red cells (which have variable blue–grey colouration (polychromasia) owing to the presence of ribosomal material) released by the marrow in an attempt to compensate for the shortened red cell survival (see Fig. 26.9).

Malaria

Malaria is caused by infection with parasites of the *Plasmodium* species. It causes anaemia through marrow suppression, splenic pooling and sequestration of red cells. There is also haemolysis because of removal and destruction of parasitized red cells, and intravascular haemolysis when the sporozoites are released from the infected red cells. There are four *Plasmodium* species, each having distinct morphological features that can be recognized by microscopic examination of a stained thick blood film in the majority of infected patients (see Fig. 26.10).

Normal white cell morphology

In health, five types of WBCs (also called leukocytes) are found in the peripheral blood (see Fig. 26.11). Their function is to protect the body against infection and other foreign invasion. This is facilitated by phagocytosis (neutrophils, eosinophils, basophils and monocytes) and humoral and cell mediated immunological activity (lymphocytes).

The most common cause of a raised WBC count (leukocytosis) is a neutrophil leukocytosis (neutrophilia), which may occur after severe exercise or tissue damage, and in bacterial infections, metabolic disorders and neoplasms of both haematological and non-haematological aetiology. The monocyte count may be raised in chronic bacterial infections and malignancies (again of both solid and haematological origin). A raised lymphocyte count (lymphocytosis) may be found in viral infections such as infectious mononucleosis and mumps, in chronic lymphocytic leukaemia and non-Hodgkin


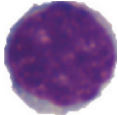

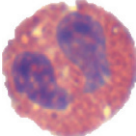
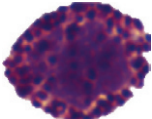
Cell type	Morphological features
	Neutrophil Deep purple staining nucleus with some chromatin clumping. Nucleus has 2–5 lobes separated by fine filaments of chromatin. Pale blue cytoplasm with an irregular outline. Numerous fine azurophilic or grey–blue granules.
	Lymphocyte Round or irregular cells with clear, pale blue cytoplasm. Nuclear chromatin varies from dense to more finely stained with occasional nucleoli present. Can be subdivided into three groups based on size and granularity: small lymphocyte large lymphocyte large granular lymphocyte.
	Monocyte Irregularly shaped cell with a large amount of cytoplasm. Cytoplasm is grey–blue and often described as having a ‘ground glass’ appearance. The cytoplasm may contain vacuoles and/or fine azurophilic granules. The nucleus makes up 70–80% of the cell and is often oval or kidney shaped.
	Eosinophil Usually has a bilobed nucleus. Pale blue cytoplasm packed with eosinophilic (red–orange) granules.
	Basophil The nucleus often has two lobes but is difficult to see as it is obscured by the large basophilic (purple) granules within the cytoplasm. The cytoplasm stains pale blue.

FIGURE 26.11 ■ White blood cells. Normal morphological features of the myeloid and lymphoid white blood cells, shown in order of prevalence in the peripheral blood.

lymphoma. Allergic disorders, parasitic infections and Hodgkin lymphoma are associated with a raised eosinophil count.

A low WBC count (leukopenia) is usually a reflection of a reduction in the concentration of neutrophils (neutropenia) or of all blood cells (pancytopenia). Causes of neutropenia include therapeutic drugs, autoimmune diseases and viral infections, whereas pancytopenia is generally caused by marrow failure.

Abnormal white cell morphology

Morphological abnormalities may be divided into those resulting from benign and malignant disorders. Benign disorders include:

- toxic changes in neutrophils in bacterial infection; neutrophils often exhibit coarser granules in the cytoplasm, cytoplasmic vacuolation and Döhle bodies (Fig. 26.12)
- neutrophil hypersegmentation in megaloblastic anaemia (Fig. 26.6)
- atypical lymphocytes in viral infection (Fig. 26.13)
- inherited disorders, for example Pelger–Huët anomaly and May–Hegglin anomaly.

Haematological malignancies

The haematological malignancies are clonal diseases originating from a single cell in the bone marrow or lymphoid tissue that has become genetically altered. They represent approximately 7% of all malignant diseases. They include acute leukaemias, chronic leukaemias, myeloproliferative neoplasms, myelodysplasia, non-Hodgkin lymphoma and multiple myeloma and its related disorders. The World Health Organization (WHO) classification of tumours of haematopoietic and lymphoid tissues is widely used to classify these malignancies. This system is based on morphologic, immunophenotypic, cytogenetic, molecular genetic and clinical features.

Acute leukaemia

Acute leukaemias have a rapid onset and progression and are invariably fatal if left untreated. They are characterized by an accumulation of early haematopoietic cells, known as blast cells (see Fig. 26.14), in the bone marrow. Morphologically, acute leukaemia is defined as the presence of >20% blast cells in the bone marrow or

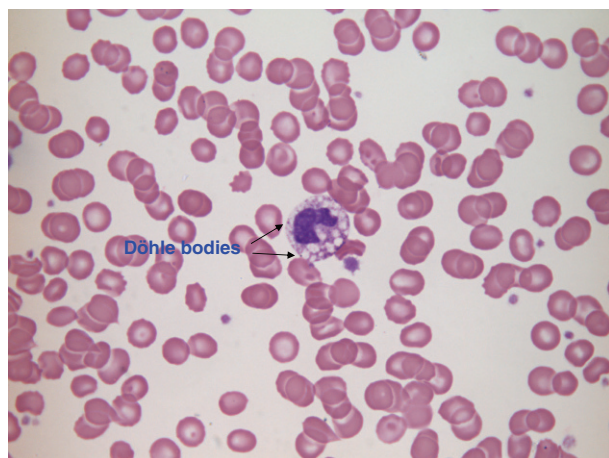


FIGURE 26.12 ■ Bacterial infection. Hb 101 g/L, WBC $4.9 \times 10^9/L$, PLT $218 \times 10^9/L$. Blood film shows toxic changes in a neutrophil: coarse granulation of the nucleus, cytoplasmic vacuolation and Döhle bodies (arrows) on the edge of the cytoplasm. Döhle bodies may also be found in myelodysplastic syndromes and acute myeloid leukaemia.

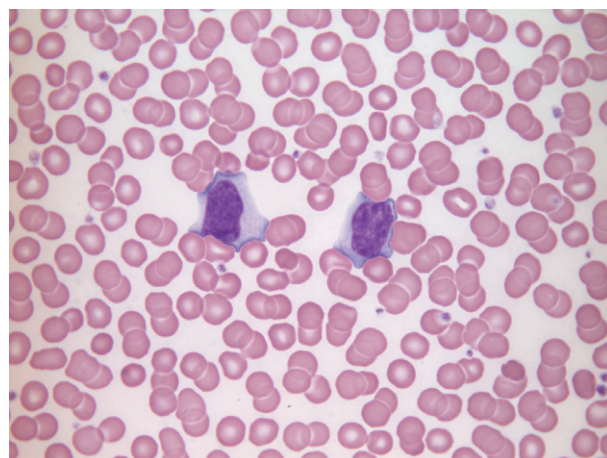
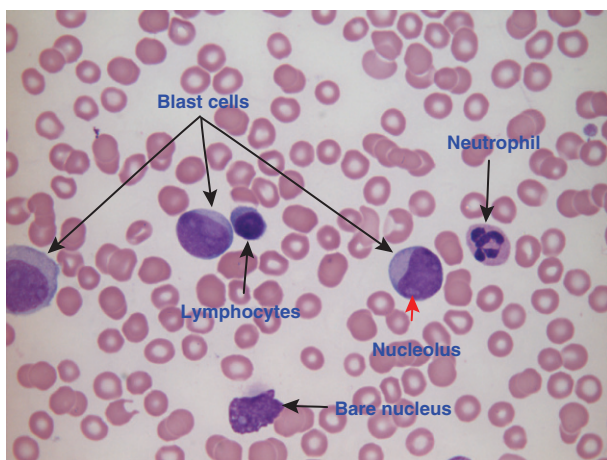
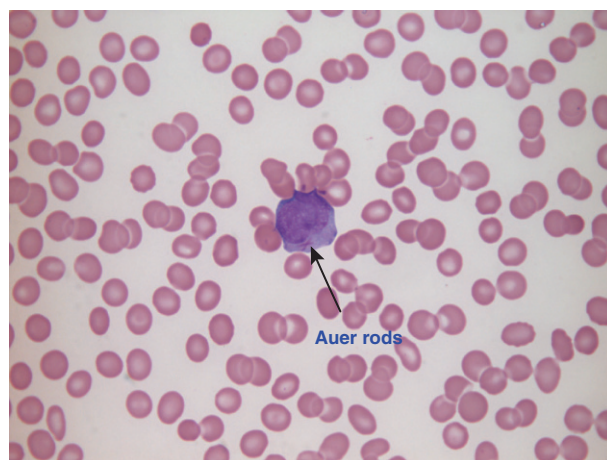


FIGURE 26.13 ■ Infectious mononucleosis. Hb 165 g/L, WBC $23.9 \times 10^9/L$, lymphocytes $13.9 \times 10^9/L$, PLT $195 \times 10^9/L$. Blood film shows two atypical lymphocytes with basophilic cytoplasm. These are T lymphocytes that react against B lymphocytes infected with EBV. The film shows a typical appearance of red cells 'wrapped' around the lymphocytes.



A



B

FIGURE 26.14 ■ Two cases of acute myeloid leukaemia (AML). (A) Hb 111 g/L, WBC $1.0 \times 10^9/L$ (3.6–11.0), blast cells $0.69 \times 10^9/L$, PLT $16 \times 10^9/L$. Blood film shows three myeloblasts containing two or more nucleoli. The film also shows a neutrophil and a lymphocyte. (B) Hb 127 g/L, WBC $30.3 \times 10^9/L$, blast cells $13.3 \times 10^9/L$, PLT $31 \times 10^9/L$. Blood film shows a myeloblast containing two Auer rods in the cytoplasm and reduced platelet numbers.

peripheral blood. The acute leukaemias are a heterogeneous group of conditions; the role of morphology in classifying them into subtypes of myeloid and lymphoid leukaemias has been superseded by the use of molecular genetics, cytogenetics and immunophenotyping. This allows classification of these complex disorders into prognostically relevant categories, and aids decision-making over the role of chemotherapy or the requirement for bone marrow transplantation after disease remission has been induced. The use of molecular genetics is becoming increasingly important in monitoring both the response of the disease to treatment and detecting patients at risk of relapse.

Chronic leukaemia

Chronic leukaemias are characterized by their slower disease progression and can broadly be divided into myeloid and lymphoid types.

Chronic myeloid leukaemia is a clonal disorder of pluripotent stem cells, as shown by the presence of the Philadelphia (Ph) chromosome, an abnormal chromosome 22, which results from translocation $t(9:22)(q34;q11)$ between chromosome 9 and 22. The fusion gene *BCR-ABL1* is also created as a result of the translocation. The white cell count is usually $>50 \times 10^9/L$ and the blood film shows all the stages of the myeloid series, from myeloblasts through to neutrophils, and an increase in basophils (see Fig. 26.15).

Chronic lymphoid leukaemias are characterized by an accumulation of mature B or T lymphocytes. Included in this group are chronic lymphocytic leukaemia, prolymphocytic leukaemia, hairy cell leukaemia and plasma cell leukaemia. Chronic lymphocytic leukaemia is the most common of the chronic lymphoid leukaemias and occurs in older subjects, typically between 60 and 80 years of age. Immunophenotyping shows the cells to be of B-lineage.

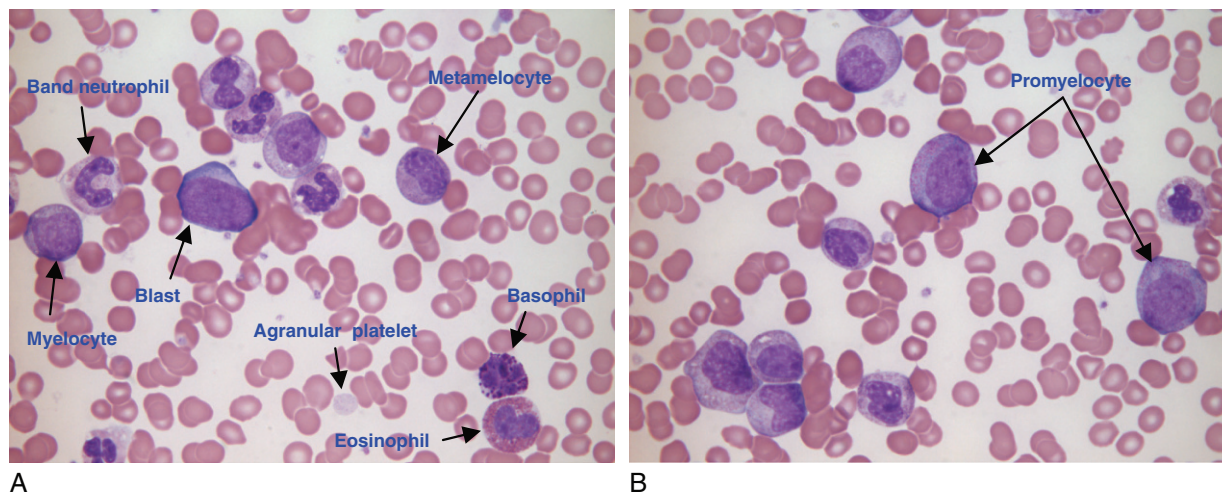


FIGURE 26.15 ■ Chronic myeloid leukaemia. Hb 101 g/L, WBC $125.6 \times 10^9/L$, PLT $338 \times 10^9/L$. (A) and (B) are photographs of the same blood film showing all stages of differentiation of the myeloid lineage. Usually this is only seen in the bone marrow.

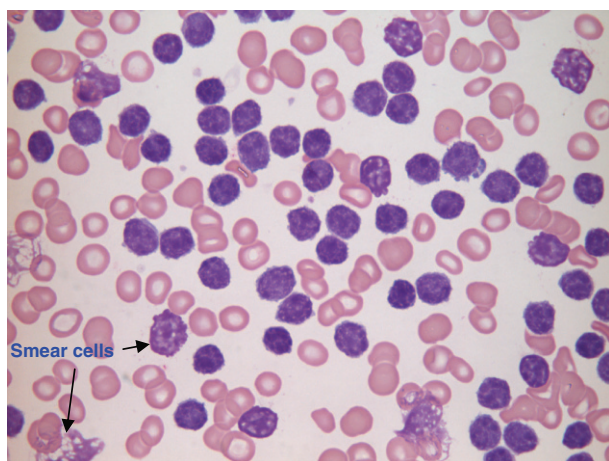


FIGURE 26.16 ■ Chronic lymphocytic leukaemia. Hb 73 g/L, WBC $950 \times 10^9/L$ (no neutrophils), PLT $85 \times 10^9/L$. Blood film shows numerous small lymphocytes and some smear cells (fragile lymphocytes damaged during spreading of blood film).

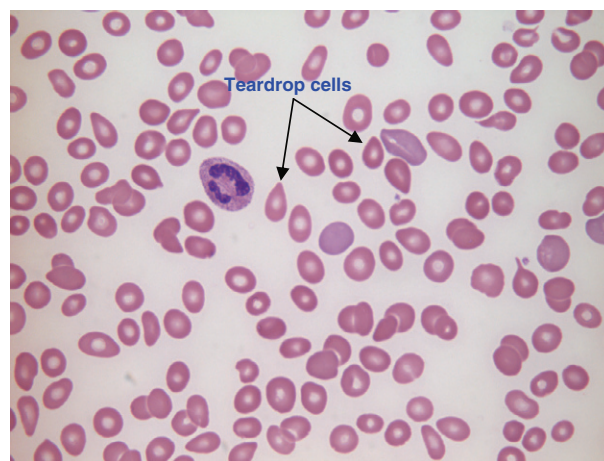


FIGURE 26.17 ■ Myelofibrosis. Hb 86 g/L, WBC $4.3 \times 10^9/L$, PLT $56 \times 10^9/L$. Blood film shows numerous teardrop cells, some polychromatic cells, one neutrophil and few platelets.

The lymphocyte count is usually $>5 \times 10^9/L$ and the blood film shows presence of predominantly small lymphocytes and smear (smudge) cells (see Fig. 26.16).

Myeloproliferative neoplasms

These are clonal disorders of haemopoietic stem cells, characterized by increased marrow proliferation of one or more myeloid cell lineages: granulocytic, erythroid, megakaryocytic and mast cells. The major myeloproliferative disorders are polycythaemia vera, essential thrombocythaemia and primary myelofibrosis (see Fig. 26.17). A single acquired mutation of the cytoplasmic kinase Janus-associated kinase 2 (JAK-2) is found in 90% of patients with polycythaemia vera and in 50% of those with essential thrombocythaemia and primary myelofibrosis.

Myelodysplasia

Myelodysplastic syndrome is a term used to describe a heterogeneous group of clonal disorders of haemopoietic stem cells. The bone marrow shows increased

proliferation but, owing to ineffective haemopoiesis, there is a reduction in all myeloid cell lineages (pancytopenia) in the peripheral blood. There is dysplasia in at least one of the cell lineages. The blood film often shows hypogranular neutrophils, pseudo-Pelger neutrophils, large hypogranular platelets and a dimorphic red cell population (see Fig. 26.18). Myelodysplastic syndrome occurs in older subjects, with a median age of 69 years. The risk of developing acute leukaemia depends upon numerous prognostic factors, including the number of blasts present in the marrow, the number of blood cell types affected and cytogenetic changes.

Non-Hodgkin lymphoma

Non-Hodgkin lymphoma refers to a heterogeneous group of clonal lymphoid tumours. The term lymphoma is generally used to describe tumours arising in lymph nodes, the spleen and other solid organs, in contrast to leukaemias, in which the bone marrow is predominantly involved, but in some cases there is

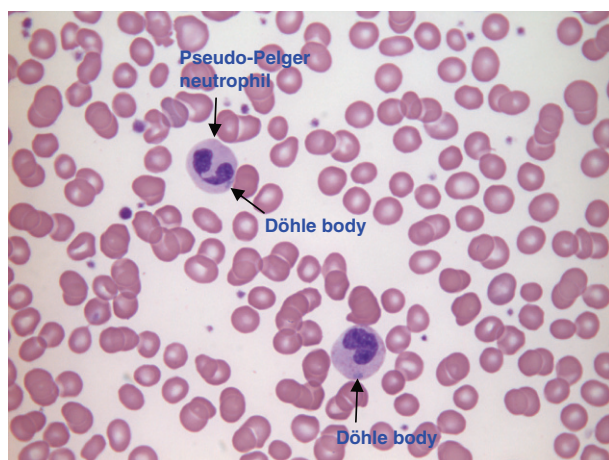


FIGURE 26.18 ■ Myelodysplasia (MDS). Hb 110 g/L, WBC $33.6 \times 10^9/L$, PLT $209 \times 10^9/L$. Blood film shows two dysplastic neutrophils with few or no cytoplasmic granules. One of the neutrophils is a pseudo-Pelger cell (bi-lobed). Both neutrophils contain Döhle bodies.

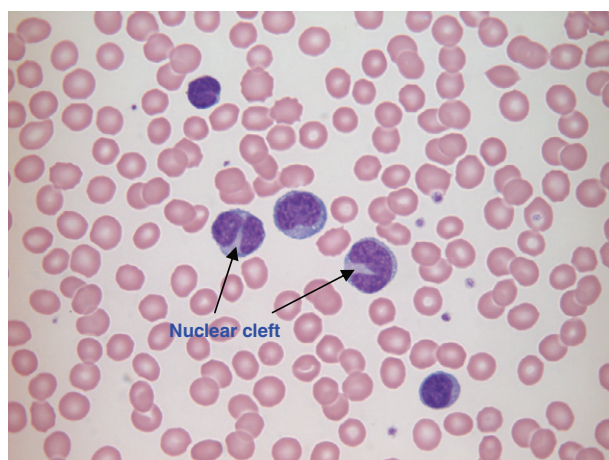


FIGURE 26.19 ■ Follicular lymphoma Hb 101 g/L, WBC $38.5 \times 10^9/L$, lymphocytes $37.7 \times 10^9/L$, PLT $85 \times 10^9/L$. Blood film shows abnormal lymphocytes, two of which have deep nuclear clefts.

no clear distinction between the two. Characteristic lymphoma cells may be seen in the blood film, for example lymphocytes with a cleaved nucleus in follicular lymphoma (see Fig. 26.19).

Multiple myeloma and related disorders are clonal malignancies of B lymphocytes. They are discussed in Chapter 30.

HAEMOSTASIS

Introduction

Haemostasis plays an essential role in preventing excess blood loss following injury. It is a complex process involving interactions between blood vessel walls, platelets and coagulation factors to form a haemostatic plug with the consequent activation of fibrinolysis, whereby the clot is gradually dissolved. Excessive activation of procoagulant mechanisms can lead to thrombosis and vascular

occlusion; therefore a delicate balance is maintained via negative and positive feedback pathways.

Following blood vessel injury, the vessel constricts to reduce blood flow, and subendothelial collagen and tissue factor is exposed. Platelets adhere to the collagen and are activated, releasing procoagulant particles and undergoing a change in membrane configuration to expose negatively charged phospholipids and coagulation factor receptors. The initial platelet plug formed is unstable but is rapidly stabilized by activation of the coagulation cascade, resulting in the formation of fibrin.

The coagulation cascade

Historically, the coagulation process was divided into separate *intrinsic* and *extrinsic* pathways, terminating in the final *common* pathway with the formation of a stable fibrin clot. It is now established that the coagulation cascade is a single series of more complex interactions.

In vivo, blood coagulation is initiated by the exposure of tissue factor (TF) following endothelial cell damage. The TF binds to factor VII, and functions as a cofactor in the activation of factor VII, with the resulting TF:FVII_a complex activating factor X and factor IX, enabling the conversion of prothrombin to thrombin. The TF:FVII_a complex is quickly inhibited by tissue factor pathway inhibitor but the thrombin already generated enables the coagulation process to continue as it feeds back to activate coagulation factors V, VIII and XI, resulting in amplification of the brief initiator process. The ultimate goal of the coagulation process is the generation of thrombin, which converts fibrinogen to fibrin. Fibrin monomers spontaneously join together via hydrogen bonds, forming insoluble fibrin polymers; these are then cross linked by factor XIII_a, changing the relatively unstable primary platelet plug into a stable clot. In vitro, however, consideration of coagulation screen results in terms of the traditional pathways allows for the simpler interpretation of the common laboratory tests (see Fig. 26.20).

Once a stable clot has been produced and bleeding has ceased, the clot is reorganized and resorbed by a process called fibrinolysis. The main enzyme responsible is plasmin, which is formed from plasminogen.

In order to restrict coagulation factor activation to the site of injury, naturally occurring inhibitors including anti-thrombin, protein C and protein S are produced.

Laboratory tests of coagulation

Laboratory tests of coagulation attempt to mimic in vitro the complex in vivo coagulation processes, but cannot do so completely. The generation of an abnormal result does not therefore necessarily mean that the patient has a bleeding diathesis, and a normal result does not completely exclude a bleeding diathesis. It is important that the test results are interpreted in the clinical context; the indiscriminate use of coagulation screens is best avoided.

The standard tests for coagulation include prothrombin time (PT), activated partial thromboplastin time (APTT), fibrinogen concentration, D-dimer concentration and

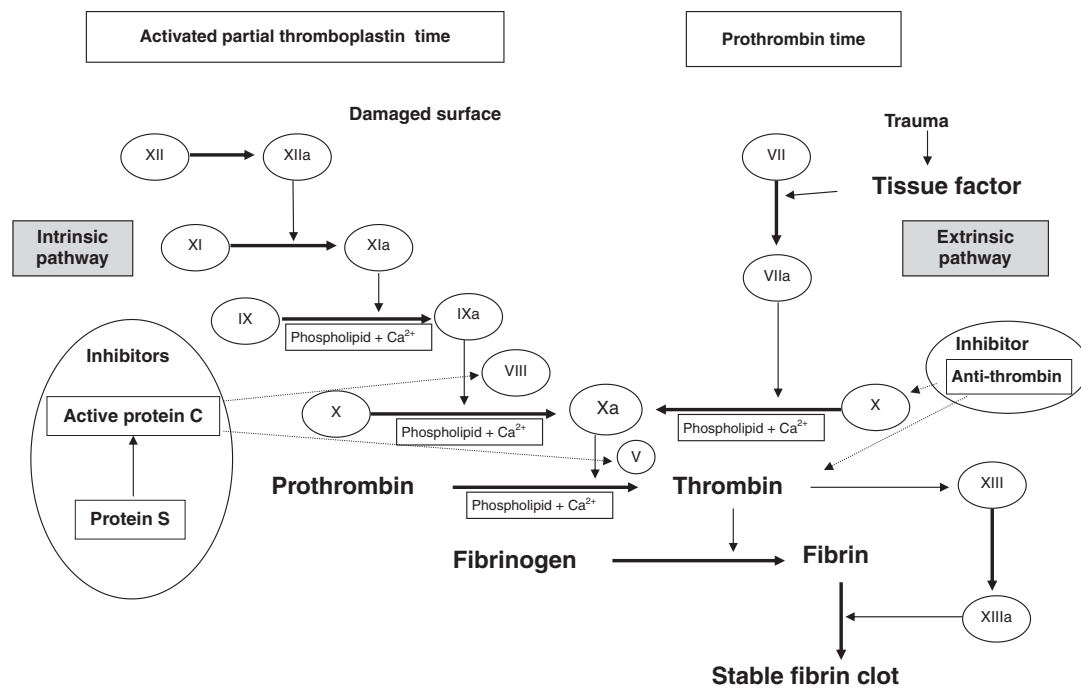


FIGURE 26.20 ■ Coagulation cascade to illustrate use of laboratory tests and the function of naturally occurring inhibitors.

thrombin time (TT). The PT measures the coagulation factors involved in the classic extrinsic and common pathways and the APTT is used to test the coagulation factors involved in the classic intrinsic and common pathways.

Prothrombin time and APTT are measured by techniques that ultimately detect the formation of fibrin clot by either mechanical or optical methods.

Prothrombin time

Plasma is mixed with thromboplastin, which contains TF, phospholipids and calcium ions, all of which are required for the activation of the extrinsic pathway. This causes a fibrin clot to form via the extrinsic and common pathways. Different thromboplastin reagents result in differing prothrombin times but internationally standardized results can be generated by correlating the thromboplastin used with the WHO reference standard. This converts the PT result to an International Normalized Ratio (INR), which is used for monitoring patients on warfarin.

Activated partial thromboplastin time

Measurement of the APTT involves induction of the formation of a fibrin clot via the intrinsic and common pathways. Plasma is first incubated with partial thromboplastin, which contains a contact activator, e.g. micronized silica or kaolin, which activates factor XII and phospholipids, triggering the intrinsic pathway. The factor XII_a in turn activates factor XI; this reaction is not significant *in vivo* but is used *in vitro* to produce factor XI_a independently of the extrinsic pathway. A second reagent containing calcium ions is then added, to enable the process to proceed to the formation of a fibrin clot.

Fibrinogen and thrombin time

Fibrinogen concentration can be estimated from the optical density changes observed when PT is measured using the optical method. A more accurate measurement can be performed using the thrombin clotting time. In this investigation, a thrombin-containing reagent is added to a dilution of the patient's plasma: the resultant clotting time is proportional to fibrinogen concentration.

D-dimer concentration

D-dimers are cross-linked D fragments: they are only produced as a result of the breakdown of cross-linked fibrin and not from the degradation of fibrinogen and soluble fibrin. Around 2–3% of plasma fibrinogen is physiologically converted to cross-linked fibrin and then degraded. Thus a small quantity of D-dimer is present in the plasma of healthy individuals. Increased amounts of cross-linked fibrin degradation products are found in states of coagulation activation and increased plasma D-dimer concentration may indicate an increased rate of fibrinolysis; this may be a consequence of the presence of a deep venous thrombosis (DVT) or pulmonary embolus (PE), or a result of disseminated intravascular coagulation (DIC). It should be noted that the main diagnostic value of D-dimer measurement lies in its negative predictive value; raised concentrations are commonly found post surgery, in pregnancy, in malignancy and in other situations associated with activation of the fibrinolytic system, resulting in many false positives.

Plasma D-dimer concentration is usually measured by latex particle enhanced immunoassay. There are also many assays that measure fibrin degradation products (FDPs) often using latex agglutination techniques.

The diagnostic value of FDPs is reduced by the fact that they may be generated from the breakdown of both fibrin and fibrinogen and thus their presence is not necessarily associated with clot formation.

Specific factor assays

Deficiency of individual clotting factors is detected by adding the patient's plasma to artificially produced plasma that is deficient in one specific factor. Measurement of PT or APTT on this mixture enables calculation of factor concentration. Results are expressed as a percentage of normal activity. An abnormal result may reflect the presence of an inhibitor antibody as well as factor deficiency. This can be identified by performing a correction study where plasma from healthy subjects is added to the test plasma, in a 50:50 ratio. If correction is less than expected, the presence of an inhibitor of coagulation should be suspected.

Interpretation of coagulation tests

If a patient presents with a history of excessive bleeding or easy bruising, the use of the simple laboratory screening tests will guide the clinician to the likely underlying cause and what further investigations are appropriate. A guide to the interpretation of the results is given in [Table 26.2](#).

Haemophilia

The most common clotting factor deficiency associated with bleeding is factor VIII deficiency, also known as haemophilia A, followed by factor IX deficiency

(haemophilia B). Inheritance of both conditions is sex-linked and can be caused by a number of different mutations. The severity of the bleeding disorder depends on the amount of active clotting factor: individuals with <1% are classified as having severe haemophilia, those with 1–5% have moderate haemophilia and those with mild haemophilia have 5–40% of normal concentrations of active clotting factor. Patients with severe haemophilia suffer from recurrent severe and spontaneous bleeds. Bleeding may occur anywhere in the body but those that are associated with the most serious complications are joints, muscles, gastrointestinal tract and the brain. Most patients with severe haemophilia require regular prophylactic supplementation with recombinant factor concentrate and additional supplementation following episodes of bleeding or to cover surgery (see [p. 513](#)). Patients may develop antibodies (or inhibitors) to the factor infused, rendering it ineffective. In these situations, recombinant activated factor VII and activated prothrombin complex concentrates (factor VIII inhibitor bypassing activity, FEIBA) can be useful in the treatment of bleeding. Patients with mild haemophilia tend to bleed only after surgery or serious trauma and can often be managed successfully with desmopressin, which releases stored factor VIII from blood vessel walls.

Disseminated intravascular coagulation

Disseminated intravascular coagulation can become catastrophic if not treated rapidly. It may be triggered by numerous disorders, such as sepsis, disseminated malignancy and amniotic fluid embolism. These lead to the formation of small blood clots in blood vessels throughout the body,

TABLE 26.2 Screening tests used in the diagnosis of coagulation disorders

Screening tests	Coagulation disorders indicated by abnormal result	Most common cause of coagulation disorder
Prothrombin time (PT)	Deficiency or inhibition of one or more of the following coagulation factors: VII, X, V, II or fibrinogen	Warfarin therapy Prematurity Disseminated intravascular coagulation (DIC)
Activated partial thromboplastin time (APTT)	Deficiency or inhibition of one or more of the following coagulation factors: XII, XI, IX VIII, X, V, II or fibrinogen	Haemophilia (factor VIII) Christmas disease (factor IX) Heparin therapy Liver disease DIC
Fibrinogen	Inflammatory or neoplastic conditions	Low concentrations: DIC Liver disease Fibrinolytic therapy High concentrations: Inflammation Neoplastic conditions
D-Dimer	Fibrin clot lysis	Pulmonary embolus Deep vein thrombosis DIC Postoperative Infection Cancer Pregnancy (raised in normal pregnancy)
Thrombin time	Deficiency or abnormality of fibrinogen or inhibition of thrombin by heparin or fibrinogen degradation products (FDPs)	Heparin therapy DIC Other thrombin inhibitors Hypofibrinogenaemia Dysfibrinogenaemia

consuming coagulation factors and platelets and resulting in abnormal bleeding, e.g. from venesection sites, gastrointestinal tract and surgical wounds. Coagulation screen results will show prolonged PT and APTT, and low fibrinogen and high D-dimer concentrations. In these circumstances, the underlying cause should be corrected as soon as possible and platelets, fresh frozen plasma (FFP) and cryoprecipitate infused if bleeding is severe (see p. 513).

BLOOD TRANSFUSION

Introduction

In the UK, blood donation is entirely voluntary and open to healthy individuals aged 17–65 years at the time of first donation. Donors undergo rigorous health assessments; individuals with lifestyle risks or a recent history of infection are excluded from donating. Each donation is of 450 mL of whole blood and is tested for human immunodeficiency virus (HIV), hepatitis B and C, and syphilis, along with determination of the Rh and ABO blood group, before it is released to be transfused into a recipient.

Hospital blood transfusion laboratories perform a range of tests required for supply of patient-compatible blood and blood products and in the investigation of haemolysis resulting from blood incompatibility (including haemolytic disease of the newborn).

Blood group antigens

The red cell membrane carries a range of surface antigens. Those that differ between individuals of the same species as a result of genetic polymorphisms (alloantigens), and can therefore elicit an immune response between donor and recipient blood, are termed the blood group antigens. There are around 30 recognized blood group systems.

ABO blood group

This blood group system is considered the most important, because of the high risk of a severe haemolytic transfusion reaction if ABO incompatible blood is transfused.

There are three antigens in the ABO blood group system, A, B and H. The *ABO* gene is located on chromosome 9 and has three alleles, *A*, *B* and *O*. The *A* allele encodes a glycosyltransferase that adds N-acetylgalactosamine to the glycoprotein H antigen that is expressed on all normal red cells. The *B* allele encodes a different glycosyltransferase that adds D-galactose. The *O* allele is a deletion that results in loss of enzyme translation and

therefore presence of unmodified H antigen. *A* and *B* are co-dominant alleles; AB individuals express both antigens.

Individuals produce antibodies to the antigen that they lack on their red cells. Landsteiner's law states that, for whichever ABO antigen is not present on the red cells, the corresponding antibody is found in the plasma (see Table 26.3). Antibodies to A and B antigen develop in the first few months of life, without exposure to the red cell antigen, as a result of environmental exposure to bacteria. They are referred to as 'naturally occurring' antibodies. ABO antibodies are usually IgM but may be IgG.

It is essential that the ABO group of transfused red cells is compatible with the ABO blood group of the patient. Transfusion of even a few millilitres of red cells into patients who have the corresponding antibody results in a severe immune reaction, which may be fatal. Cells of blood group O are compatible with patients of all ABO blood groups so individuals of group O can be referred to as *universal donors*. Conversely, individuals of blood group AB can receive blood from any of the other groups and are referred to as *universal recipients* (see Table 26.3).

Rh blood group

The Rh (formerly rhesus) blood group consists of over 45 antigens, of which D, C, E, c and e antigen are usually typed in donor blood. All patient samples for blood grouping are routinely typed for their RhD status because the D antigen is clinically the most important after the ABO antigens. About 85% of the population express the D antigen and are called RhD positive; the remainder are termed RhD negative. The D antigen is highly immunogenic so individuals who are RhD negative but receive RhD positive blood are likely to produce anti-D antibodies. Anti-D has the potential to cause a severe haemolytic transfusion reaction and cause haemolytic disease of the newborn (HDN) (see below).

Expression of the D antigen is variable. Some patients express only small amounts of the D antigen; their RhD status is termed weak D. Other patients express only part of the D antigen; their RhD status is described as partial D. Partial RhD individuals can produce antibody to the part of the D antigen, which they lack, so they should receive RhD negative red cells if they require transfusion.

Other important blood groups and antibodies

Kell, Kidd, Duffy and MNS blood groups are clinically important because antibodies to antigens of these blood groups can cause either transfusion reactions or HDN or both. They are included routinely in antibody identification panels.

TABLE 26.3 ABO antigen, antibodies, genotype and compatibility

ABO group	Antigen on red cell	Antibody in plasma	Genotype	Compatible ABO group
A	A	Anti-B	AA or AO	A, O
B	B	Anti-A	BB or BO	B, O
O	None	Anti-A,B	OO	O
AB	A and B	None	AB	AB, A, B, O

The Kell antibody (Anti-K) is the most commonly found antibody after those to the ABO and Rh blood group antigens. A third of all non-ABO, non-Rh allo-antibodies are anti-K. It is good practice to give Kell-negative blood to patients who already have antibodies to other blood group antigens. Women of child-bearing age who require blood transfusion should also be transfused with Kell negative blood of the appropriate ABO and Rh grouping.

The Kidd antibodies (Anti-Jk(a) and Anti-Jk(b)) form another important antibody system that can cause severe acute and delayed transfusion reactions. Anti-Jk(a) is the most commonly implicated antibody in delayed transfusion reactions. Anti-Jk(a) may fall to low or undetectable concentrations in the patient's plasma over time so can be difficult to detect.

Laboratory transfusion tests

Blood grouping and antibody screen

Commonly referred to as 'group and save', this is the most requested test and comprises two parts: ABO and RhD grouping, and an antibody screen.

ABO and D grouping Antisera are mixed with the patient's red cells; this is known as the forward group. In ABO grouping, the patient's plasma is also mixed with known A and B cells: the reverse group. The positive and negative patterns of agglutination determine the patient's ABO and RhD group (see Table 26.4). This can be performed manually or by automated methods but the principles remain the same.

Antibody screening This procedure is designed to detect clinically significant alloantibodies to blood group systems other than ABO. There are two stages to an antibody screen: sensitization and haemagglutination.

The patient's plasma is incubated at 37°C with two or three red cell samples of known blood group phenotype. These samples are always group O to avoid reactions that would result from ABO incompatibility. The antibody-antigen complex formation that occurs is termed sensitization. In the second stage, anti-human globulin (AHG) reagent is added, to encourage cross linking between sensitized red cells, as most antibodies are of the IgG class so cannot directly form cross links. If agglutination occurs, further tests are required to identify the antibody or antibodies present. This procedure is known as the indirect antiglobulin test (IAT).

The patient's sample is retained and stored in accordance with national guidelines. It can subsequently be used in order to crossmatch blood for that patient, as described later. It is essential that, if a patient has recently been transfused, they are re-tested prior to any further transfusion in order to detect any new antibodies they may have developed in response to the previous transfusion.

Antibody identification panels

Antibodies that need to be identified are those that can cause transfusion reactions or HDN. Following a positive antibody screen, further tests are done to identify the antibody using the same theory as for antibody screening. The antibody panel usually contains at least ten different red cell reagents of known phenotype. There are usually two sets of red cell reagents in one panel set, one of which has been treated with an enzyme that cleaves off part of the red cell membrane, as a result of which some antigens are lost from the red cell membrane (MNS and Duffy systems) and others are exposed in order to produce stronger antibody binding (the Rh system). The use of both sets of cells is important as the differences in reaction for different blood group systems can help determine antibody identity, particularly when multiple antibodies are present.

Crossmatching (compatibility testing)

Crossmatching is designed to ensure compatibility of donor units with the recipient. The procedure can be either non-serological or serological. There are three main types of crossmatch:

- electronic crossmatch or blood issue (non-serological)
- immediate spin crossmatch (serological)
- full crossmatch (serological).

Electronic crossmatch is the quickest to perform, enabling issue of blood within five minutes. A prerequisite is that a group and antibody screen has already been performed for that patient. In this non-serological process, the laboratory computer system checks for ABO and RhD compatibility. For a sample to be suitable for electronic crossmatching, a number of requirements must be met. There are some minor differences between guidelines issued by the British Committee for Standards in Haematology (BCSH) and the American Association of Blood Banks (AABB), but the following elements are generally considered essential in the UK:

TABLE 26.4 A selection of ABO and D grouping results

Known reagent	Anti-A	Anti-B	Anti-A,B	Anti-D	Anti-D	A cells	B cells	
Patients cells/plasma	Cells	Cells	Cells	Cells	Cells	Plasma	Plasma	
	-	-	-	+	+	+	+	O Pos
	+	-	+	+	+	-	+	A Pos
	-	+	+	-	-	+	-	B Neg
	+	+	+	-	-	-	-	AB Neg
	Forward group				Reverse group			

- the ABO and RhD grouping procedure must have been performed on a fully automated grouping machine
- results must have been transferred to the laboratory computer system electronically, with no manual editing of results
- there must be two concordant ABO and RhD records on file for that patient
- no clinically significant antibodies have been detected presently or historically.

Transfusion laboratories may also adhere to additional locally determined rules, for example only donor units that have been electronically entered onto the computer system (via barcode scanning) being considered suitable for electronic issue. Not all laboratories perform electronic crossmatch, as some may not be equipped with automated analysers or do not have a laboratory computer system that is capable of processing the required algorithm.

Immediate spin crossmatch is designed principally to detect ABO incompatibility. It may be used in an emergency when there is insufficient time to perform a full antibody screen. It may also be performed when the criteria for electronic issue are not fully met, for example if a sample has been run on an automated analyser that has not been able to interpret the group correctly. This could be caused by weak expression of anti-A, anti-B or RhD. It is also important that there is no record of the patient having clinically significant antibodies presently or historically. This is a manual technique performed in tubes: the patient's plasma and donor red cells are incubated and then centrifuged to check for haemagglutination or haemolysis that would indicate ABO compatibility. Using this technique, it takes 10–15 min for blood to be issued from the laboratory.

Full crossmatch is performed if time or equipment availability do not permit a sample to be processed on a blood grouping analyser, or following automatic processing if the patient is known to have clinically significant antibodies presently or historically. Some laboratories may not have adopted the immediate spin or the electronic crossmatch and therefore continue to use the full crossmatch routinely. In this procedure, the patient's plasma is incubated with donor units and checked for compatibility using the IAT technique described in the previous section. Using this technique, it takes 45 min for blood to be issued from the laboratory.

Investigation of suspected transfusion reaction

Serious or life-threatening acute reactions are very rare. Acute haemolytic transfusion reactions are most commonly a consequence of errors in taking or labelling the sample, collecting wrong blood from the blood bank, inadequate checking of the product at the bedside or laboratory errors. Symptoms, such as the patient having a feeling of impending doom, agitation, flushing or pain at the venepuncture site or in the abdomen, flank or chest can occur after transfusion of only a few drops of blood.

Clinical signs include fever and hypotension; generalized oozing from wounds or puncture sites can occur and there may be haemoglobinaemia or haemoglobinuria. Life-threatening transfusion reactions can also occur with platelet and FFP transfusions, for example owing to bacterial contamination or transfusion-related acute lung injury.

Following a suspected transfusion reaction, urgent blood tests (full blood count, assessment of kidney and liver function), along with assessment of the urine for haemoglobinuria, should be undertaken. The donor units and giving set should be returned to the laboratory, along with a fresh blood sample. The blood sample and donor units along with the original 'group and save' sample are then tested retrospectively for compatibility. At this point, it is also important to confirm the blood group of any units transfused and send the giving set and donor packs for culturing. Blood cultures should be taken if the patient has a persistent fever, and a coagulation screen performed to exclude developing DIC if the reaction is severe.

Haemolytic disease of the newborn

Haemolytic disease of the newborn (HDN) is a condition in which the neonatal red cells have a shortened lifespan because antibodies of maternal origin have crossed the placenta and coated them. These cells are then removed by the immune system. The mother will have produced the antibodies following fetal–maternal haemorrhage. It is not uncommon in pregnancy for small 'silent' bleeds to occur that are too small for detection. Antibodies that cause HDN are of the IgG class. The most common cause of HDN is ABO incompatibility, in which cases the haemolysis is usually mild. More severe cases of HDN can be caused by anti-D, anti-c and anti-K. Anti-K suppresses neonatal red cell production as the K antigen is one of the first antigens to be expressed on the red cells during red cell production.

Standard UK antenatal practice is that samples are taken for blood group and antibody screening when the patient is booked for antenatal care (8–12 weeks of gestation) and again at 28 weeks. If the mother is RhD negative, samples are also sent at delivery. Expectant mothers who have produced antibodies during pregnancy require close monitoring. Depending on when the antibody is first detected, BCSH guidelines state that grouping and antibody screening should be performed and antibody titre determined every four weeks up to the 28th week of pregnancy and then every two weeks thereafter. The antibody titre can be used as a guide to the potential severity of HDN, with the exception of anti-K for which there is no quantitative test and little correlation between concentration and the severity of HDN.

In the UK, the incidence of HDN caused by anti-D has been reduced following the implementation of national anti-D prophylaxis programmes for antenatal women. Pregnant women who are RhD-negative are given prophylactic anti-D sufficient to deal with any silent bleeds. Routine prophylaxis can either be given as two doses of 500 IU anti-D administered at 28 weeks and then 34 weeks of gestation or a single dose of

1500 IU administered at 28 weeks. At delivery, an RhD negative mother is given more anti-D if the baby is RhD positive. The amount they receive depends upon the results of a Kleihauer test, which is a slide test for the number of fetal cells in the maternal circulation (see Fig. 26.21). If any sensitizing events occur during the pregnancy, it will be necessary to administer a top-up dose of anti-D.

Blood products

Red cells

Red cells are the most common blood product issued by a transfusion laboratory. Indications for red cell transfusion include active bleeding, acute anaemia caused by trauma or surgery, and chronic anaemia secondary to malignancy. For surgery in which blood loss is expected, red cells are cross-matched in advance in accordance with a maximum surgical blood ordering schedule (a schedule of the number of red cell units routinely crossmatched for each type of elective surgical procedure). Red cell concentrates have a maximum shelf-life of 35 days and must be stored at $4 \pm 2^\circ\text{C}$.

Transfusion of emergency group O RhD negative blood is an important option in life-threatening situations when there is insufficient time to provide crossmatched blood. These units must be easily available especially to the emergency department and theatres. The criteria for the selection of donor units should be determined locally, for example negative for the C and E Rh antigens and the Kell antigen, low titre of ABO antibodies and cytomegalovirus negative.

Platelets

Platelet transfusions are indicated for the prevention and treatment of haemorrhage in patients with thrombocytopenia or defective platelet function. They are, however, contraindicated in post transfusion purpura, thrombotic thrombocytopenia purpura and heparin induced thrombocytopenia as they would increase the risk of thrombosis.

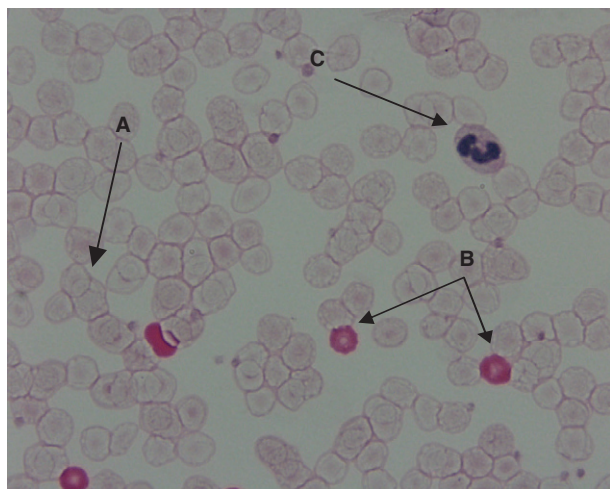


FIGURE 26.21 ■ Kleihauer film. There is a good distinction between the three cell types: maternal cells (A), fetal cells (B) and white cells (C). The presence of RhD positive fetal cells is confirmed by flow cytometry.

Platelet concentrates can be produced by pooling together platelets from whole blood donations or by obtaining platelets by apheresis. Platelets must be stored at $22 \pm 2^\circ\text{C}$ and have a shelf-life of up to seven days with continuous gentle agitation.

Fresh frozen plasma

Fresh frozen plasma (FFP) contains all clotting factors found in plasma. Indications for its use are limited but it is most often used in patients with massive bleeding where clotting factors are being consumed or are diluted owing to the administration of large volumes of red cells or crystalloids. It must be stored at $< -25^\circ\text{C}$ and for a maximum of 24 months. It is most effective if used within four hours of thawing but can be stored at 4°C post thawing for up to 24h.

Cryoprecipitate

Cryoprecipitate is a plasma fraction rich in fibronectin, fibrinogen, factor VIII and von Willebrand factor. It is used particularly in patients who have low plasma fibrinogen concentrations such as during massive transfusions and in DIC. Like FFP, it must be stored at $< -25^\circ\text{C}$ and for a maximum of 24 months.

Factor concentrates

Factor concentrates can be either fractionated or recombinant. Fractionated products are produced by pooling multiple donations, which increases the risk of viral transmission to the recipient, making recombinant products the preferred option. Recombinant factor concentrates are used in patients with inherited deficiencies of clotting factors, either prophylactically or during bleeding episodes (see p. 509). Prothrombin complex concentrate, rich in factors II, VII, IX and X, is used for the emergency correction of warfarin-induced bleeding. Factor VII concentrate, although only licensed for use in inherited or acquired haemophilia with significant inhibitors or Glanzmann thrombasthenia (a platelet function disorder), is being increasingly used if life-threatening bleeding persists despite optimization of laboratory measured clotting and platelet parameters.

Risks of transfusion

The incidence of adverse events caused by blood transfusion is low, although the consequences are potentially severe. The main adverse risk is that of receiving the incorrect blood component. Strict procedures are in place to ensure vein to vein traceability of all blood products and to reduce the risks of misidentification of either the patient or donor. The very rare but fatal complication of transfusion-associated graft versus host disease, whereby immunocompetent donor lymphocytes are infused into an immunocompromised patient (from bone marrow transplantation or other defects in T cell immunity), is prevented by the irradiation of blood products to inactivate any residual white cells. The risk of transmitting infections is very small provided that appropriate screening and selection of donors is in place.

Regulations

In the UK, blood transfusion laboratories have to comply with a range of specific regulations and standards of practice, which are monitored by the Medicines and Healthcare Regulatory Authority (MHRA). The key standards are incorporated into UK law, making infringement a criminal offence. Adverse incidents associated with blood transfusion must be reported to the MHRA.

CONCLUSION

- The full blood count is the most requested test profile in haematology and is a valuable first line screening test from which further investigations can be generated.
- Blood film examination is an essential part of a haematological investigation, as valuable information can be obtained from cell morphology.
- Prothrombin time, activated partial thromboplastin time, thrombin time, D-dimer and fibrinogen measurements are useful screening tests for coagulation disorders but should be interpreted within the clinical context.
- The blood transfusion laboratory performs a vital role in the provision of safe and effective blood product support.
- Antibody identification is an essential procedure in the prevention of transfusion reactions and haemolytic disease of the newborn.

ACKNOWLEDGEMENTS

We would like to thank our colleagues Rebecca Pleasance, Gemma Greenall and Colin Greig, who critically reviewed the material in this chapter and made many helpful suggestions. Rebecca Pleasance also provided the photographs and some of the text used in Figures 26.4 and 26.11.

Further reading

- Bain BJ. *Blood cells*. 4th revised ed. Chichester: Wiley-Blackwell; 2006.
- Bain BJ, Bates I, Laffan MA et al. editors. *Dacie and Lewis practical haematology*. 11th ed. Edinburgh: Churchill Livingstone; 2012.
- British Committee for Standards in Haematology Blood Transfusion Task Force. Guidelines for pre-transfusion compatibility procedures in blood transfusion laboratories. 2012a. www.bcshguidelines.com/documents/Compat_Guideline_for_submission_to_TTF_011012.pdf [Accessed 23 January 2013].
- British Committee for Standards in Haematology Blood Transfusion Task Force. Guidelines on the investigation and management of acute transfusion reactions; 2012b. www.bcshguidelines.com/documents/ATR_final_version_to_pdf [Accessed 23 January 2013].
- Daniels G. *Human blood groups*. 3rd ed. Chichester: Wiley-Blackwell; 2013.
- Hoffbrand AV, Moss P, editors. *Essential haematology*. 6th ed. Oxford: Wiley-Blackwell; 2011.
- Key N, Makris M, O'Shaughnessy D et al. editors. *Practical hemostasis and thrombosis*. Chichester: Wiley-Blackwell; 2009.
- Klein HG, Anstee DJ, editors. *Mollison's blood transfusion in clinical medicine*. 12th ed. Chichester: Wiley-Blackwell; 2011.
- Murphy MF, Pamphilon DH, editors. *Practical transfusion medicine*. 3rd ed. Chichester: Wiley-Blackwell; 2009.
- Provan D, Singer CRJ, Baglin T et al. *Oxford handbook of clinical haematology*. In: 3rd ed. Oxford: Oxford University Press; 2009.
- Swerdlow SH, Campo E, Harris NL et al. editors. *WHO classification of tumours of haematopoietic and lymphoid tissues*. 4th ed. Geneva: WHO; 2008.

Biochemical aspects of anaemia

Rebecca Frewin

CHAPTER OUTLINE

INTRODUCTION 515

THE FORMATION, STRUCTURE AND FUNCTION OF THE NORMAL RED CELL 515

Formation 515

Structure of the red cell 516

Function of the red cell 517

ANAEMIA 517

ANAEMIAS ASSOCIATED WITH A REDUCTION IN RED CELL PRODUCTION 517

Iron deficiency anaemia 517

The megaloblastic anaemias resulting from vitamin B₁₂ and folate deficiency 520

Anaemias due to reduction in red cell production: inherited causes 524

Anaemias associated with reduction in red cell production: acquired causes 524

ANAEMIAS ASSOCIATED WITH INCREASED RED CELL LOSS 524

Bleeding 524

Haemolysis 524

Causes of haemolytic anaemias 525

Inherited haemolytic anaemia 525

Acquired haemolytic anaemias 528

DIAGNOSIS OF HAEMOLYSIS 529

Clinical evidence 529

Laboratory investigations 529

CONCLUSION 532

INTRODUCTION

Blood is often considered a liquid organ that acts as the transport system of the body. It is composed of a liquid (plasma) in which cells are suspended. There are three types of cells. Red cells are the most numerous; they play a vital role in the provision of oxygen to the organs and peripheral tissues. The white cells comprise the phagocytes, granulocytes and monocytes, which are mobilized in the blood to sites of infection and inflammation where they pass into the tissues to engulf and destroy bacteria and other objects deemed foreign by the body, and the lymphocytes which, via the secretion of antibodies and cytokines, coordinate defence against acute infection and maintain long term immune surveillance. Finally, the platelets, in conjunction with coagulation proteins in the plasma, allow prompt and effective haemostasis in the event of injury.

This chapter will focus on the structure and function of the red cells, or erythrocytes, and the pathophysiological mechanisms underlying the disorders of red cells that result in anaemia.

THE FORMATION, STRUCTURE AND FUNCTION OF THE NORMAL RED CELL

Formation

The first primitive blood cells develop in the embryo during the third week after conception and are seen in

the yolk sac. These cells migrate to the liver and spleen, which become the major sites of blood formation (erythropoiesis) until after birth. From approximately five months' gestation, erythropoiesis begins to occur in bone marrow. This is called medullary erythropoiesis (as opposed to extramedullary erythropoiesis, when cells are produced in other sites such as the liver and spleen). By birth, medullary erythropoiesis is occurring in the marrow cavity of nearly every bone in the body. With maturity, many of the marrow cavities become replaced with fat and, by adulthood, erythropoiesis is limited to the axial skeleton (sternum, vertebrae and pelvis only). The other sites retain the ability to produce blood cells if marrow function is compromised by various pathological processes.

Red cells survive for approximately 120 days and therefore the bone marrow needs to generate new haemopoietic cells continuously. It is estimated that the average adult requires synthesis of 10^{11} red cells per day. Red cells originate from pluripotent uncommitted stem cells that are capable of producing any type of blood cell. The earliest recognizable committed progenitor for erythroid cells is the colony forming unit – granulocyte, erythroid, megakaryocyte, macrophage (CFU-GEMM). This has a limited self-renewal capacity and may mature into the various haemopoietic lineages depending on various stimuli within the bone marrow microenvironment. The first unique erythroid progenitor is the burst

forming unit – erythroid (BFU-E), from which further progeny develop into recognizable erythroid cells (see Fig. 27.1). Maturation of the erythroid cells involves primarily haemoglobin synthesis, which is vital for the subsequent function of the cell, and also condensation and eventual extrusion of the nucleus, removing the ability of the cell to proliferate. This process takes approximately seven days and results in the release of a reticulocyte, an immature red cell slightly larger than the mature red cell with a blue tinged cytoplasm owing to residual nuclear material. The reticulocyte then travels from the marrow to the spleen where the residual nuclear remnants are removed, producing the mature red cell.

Thus, effective erythropoiesis requires a delicate balance between maintaining a pluripotent stem cell pool and allowing the development and terminal differentiation of the progeny to mature erythroid cells. The exact process by which this is achieved has not been elucidated but involves many transcription factors such as runt-related transcription factor 1 (RUNX1), acute myeloid leukaemia 1 (AML-1) protein, TEL oncogene (ETV6) and mixed lineage leukaemia (MLL) protein orchestrating, either alone or in conjunction with others, the intricate signalling processes required to produce mature erythroid cells. The dysregulation of these transcription factors, often via genetic mutations, results in a range of haematological malignancies that afford scientists a means of studying the role of these

factors in normal erythropoiesis. GATA-1, with its co-factor FOG-1 (friend of GATA-1), is one of the most important transcription factors involved in the terminal differentiation of the erythroid cell. Its name is derived from its ability to bind to DNA sites with the consensus sequence (AT)GATA(AG) within the promoter regions of many of the genes encoding red cell membrane proteins and enzymes.

Erythropoiesis is the production of mature, haemoglobin-rich, red cells that carry oxygen to the tissues of the body. It is therefore not surprising that hypoxia is a major driver of erythropoiesis via the induction of hypoxia inducible factor (HIF), a transcription factor pivotal to coordinating the body's response to hypoxia. This binds to hypoxia response elements (HREs), resulting in the activation of genes for proteins that play a vital role in oxygen delivery, such as vascular endothelial growth factor (leading to new blood vessel formation) and erythropoietin (a hormone which stimulates erythropoiesis).

Structure of the red cell

Red cells have a biconcave disc shape and a membrane structure that gives them particular strength and deformability to allow their passage through capillary networks, allowing the effective oxygenation of the tissues. The membrane consists of a phospholipid bilayer, with charged hydrophilic phosphatidyl groups forming the outer and inner surfaces, and a hydrophobic interior.

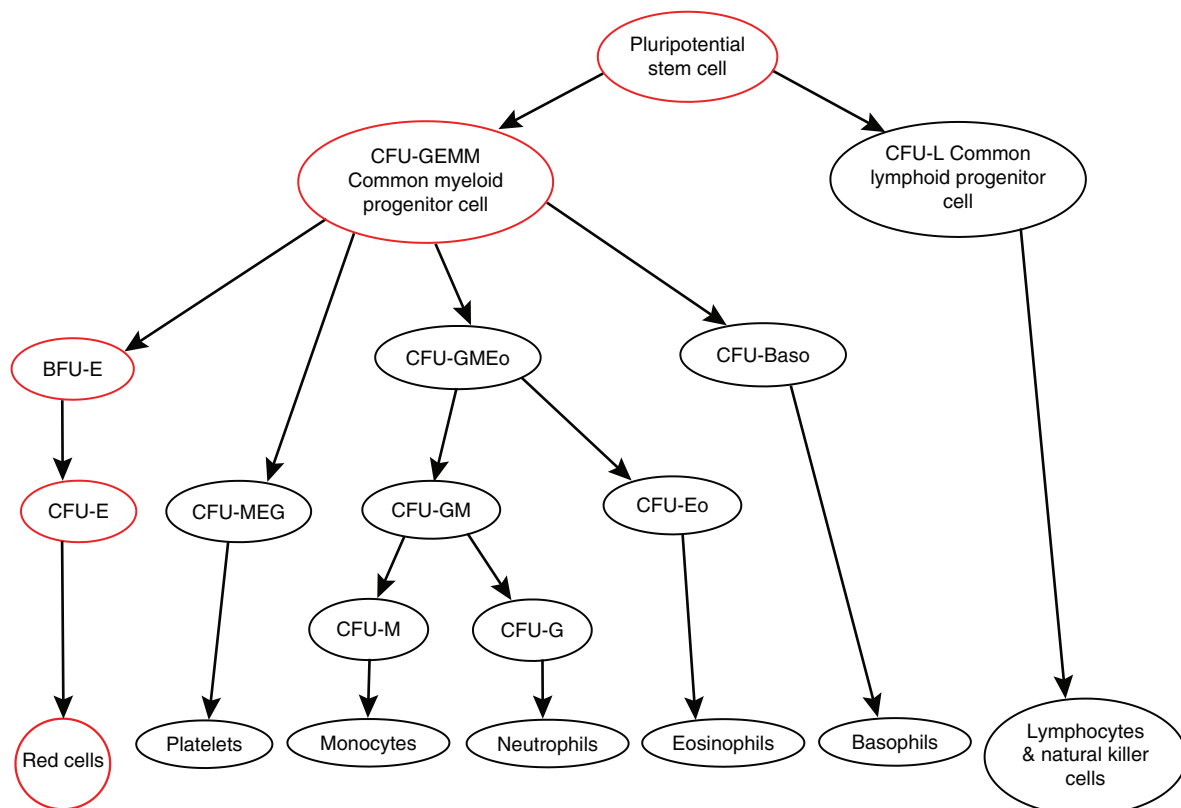


FIGURE 27.1 ■ Schematic representation of haemopoiesis. BFU, burst forming unit; CFU, colony forming unit; E, erythroid; Eo, eosinophil; GEMM, granulocyte, erythroid, monocyte, megakaryocyte; GM, granulocyte, monocyte; L, lymphocyte; M, monocyte; MEG, megakaryocyte.

It is stabilized by integral membrane proteins that span the lipid bilayer, providing numerous functions, including the transmembrane passage of ions and interaction with the proteins of the cytoskeleton below. The cytoskeletal protein spectrin forms tetramers that aggregate into a hexameric scaffolding that lies on the cytoplasmic surface of the cell membrane to preserve its shape and structure. The ankyrins anchor membrane proteins to spectrin and maintains correct orientation of membrane ion channels. Abnormalities in any of these proteins result in an alteration of shape and flexibility of the red cell with a consequent reduction in survival. This is the pathological basis of the inherited haemolytic anaemias, of which the commonest is hereditary spherocytosis.

Function of the red cell

The most important function of the red cells is to transport oxygen from the lungs to the tissues and to return the carbon dioxide they produce to the lungs. Normal adult haemoglobin consists of two α globin chains and two β globin chains; a haem prosthetic group is bound to each chain. Haem comprises a tetrapyrrole IX α ring bound to a single Fe^{2+} ion. The binding of a single molecule of oxygen to a haem group results in a change in conformation of the haemoglobin molecule that promotes further oxygen binding to the other subunits. These allosteric interactions allow for the efficient uptake and release of oxygen. In order to achieve this, the red cell requires a source of both ATP and reducing power. ATP generation is vital to maintain the deformability of the membrane and to regulate water and ion exchange. The reducing power of the red cell is required both to protect against the oxidation of the lipid bilayer and to reduce the methaemoglobin that is formed when ferrous iron (Fe^{2+}) is oxidized to its ferric form (Fe^{3+}) back to functional deoxyhaemoglobin. Anaerobic glycolysis is the main pathway responsible for supplying the cell with both ATP and nicotinamide adenine dinucleotide (reduced) (NADH), a cofactor for methaemoglobin reductase, the enzyme that catalyses the reduction of methaemoglobin to functional haemoglobin (see Fig. 27.2).

The pentose phosphate pathway produces nicotinamide adenine dinucleotide phosphate (reduced) (NADPH) and pentose sugars. The latter are returned to the mainstream of glycolysis if necessary. Reduced nicotinamide adenine dinucleotide phosphate is essential for keeping glutathione in the reduced state, thus enabling it to perform its antioxidant functions. These consist mainly of eliminating hydrogen peroxide and free radicals and thus preventing oxidation of the red cell membrane or haemoglobin. The Rapoport–Luebering shunt is responsible for producing 2,3-diphosphoglycerate (2,3-DPG) from the 1,3-diphosphoglycerate that is produced in anaerobic glycolysis. 2,3-Diphosphoglycerate modulates the oxygen affinity of haemoglobin (see Chapter 5).

Various non-glycolytic enzymes contribute to red cell metabolism in a complementary fashion. Some of these are involved in glutathione metabolism as shown in Figure 27.2. Others catalyse nucleotide metabolism

and participate in adenylate salvage pathways. During maturation of the reticulocyte, the ribosomal DNA is catabolized to its constituent nucleotides by the enzyme ribonuclease. The resulting nucleotides of adenine are useful to the cell and are processed through appropriate salvage pathways. Nucleotides of pyrimidine, however, can be harmful and must be eliminated. This is achieved through their dephosphorylation by a specific enzyme, pyrimidine 5'-nucleotidase, following which the resultant nucleosides leave the cell by passive diffusion.

As the red cell matures, it extrudes its nucleus and mitochondria, and becomes incapable of protein synthesis. Thus, none of the enzymes contained within the red cell can be replaced and the decline in metabolic activity results in a more rigid and fragile cell membrane. This becomes damaged during its passage through the capillaries and the red cell is eventually removed by cells of the reticuloendothelial system, particularly the macrophages of the spleen.

ANAEMIA

Anaemia is defined as a blood haemoglobin concentration below that which would be expected for a healthy individual of that age or sex. Regardless of its cause, the symptoms of anaemia are similar, reflecting the effects of poor oxygen delivery to tissues. These symptoms may often be vague and include tiredness, headaches and breathlessness.

Anaemia may be classified on the basis of the size of the red cells (microcytic, normocytic and macrocytic). This chapter, however, will discuss anaemias using a classification based on kinetic aspects of red cell production as follows:

- reduction in red cell production
 - nutritional deficiencies
 - inherited causes
 - acquired causes
- increased red cell loss
 - bleeding
 - haemolysis (inherited or acquired).

ANAEMIAS ASSOCIATED WITH A REDUCTION IN RED CELL PRODUCTION

Iron deficiency anaemia

Iron deficiency anaemia is one of the commonest causes of anaemia worldwide and has significant socioeconomic consequences. Other haematinics essential for adequate erythropoiesis are vitamin B_{12} and folate.

Iron physiology

Iron plays a vital role in many metabolic processes. It is an essential component of both haemoglobin and myoglobin and also plays an important part in electron transfer and the generation of energy (e.g. in cytochrome *c* oxidase and catalase). Non-haem iron is also important in DNA synthesis as it is essential for

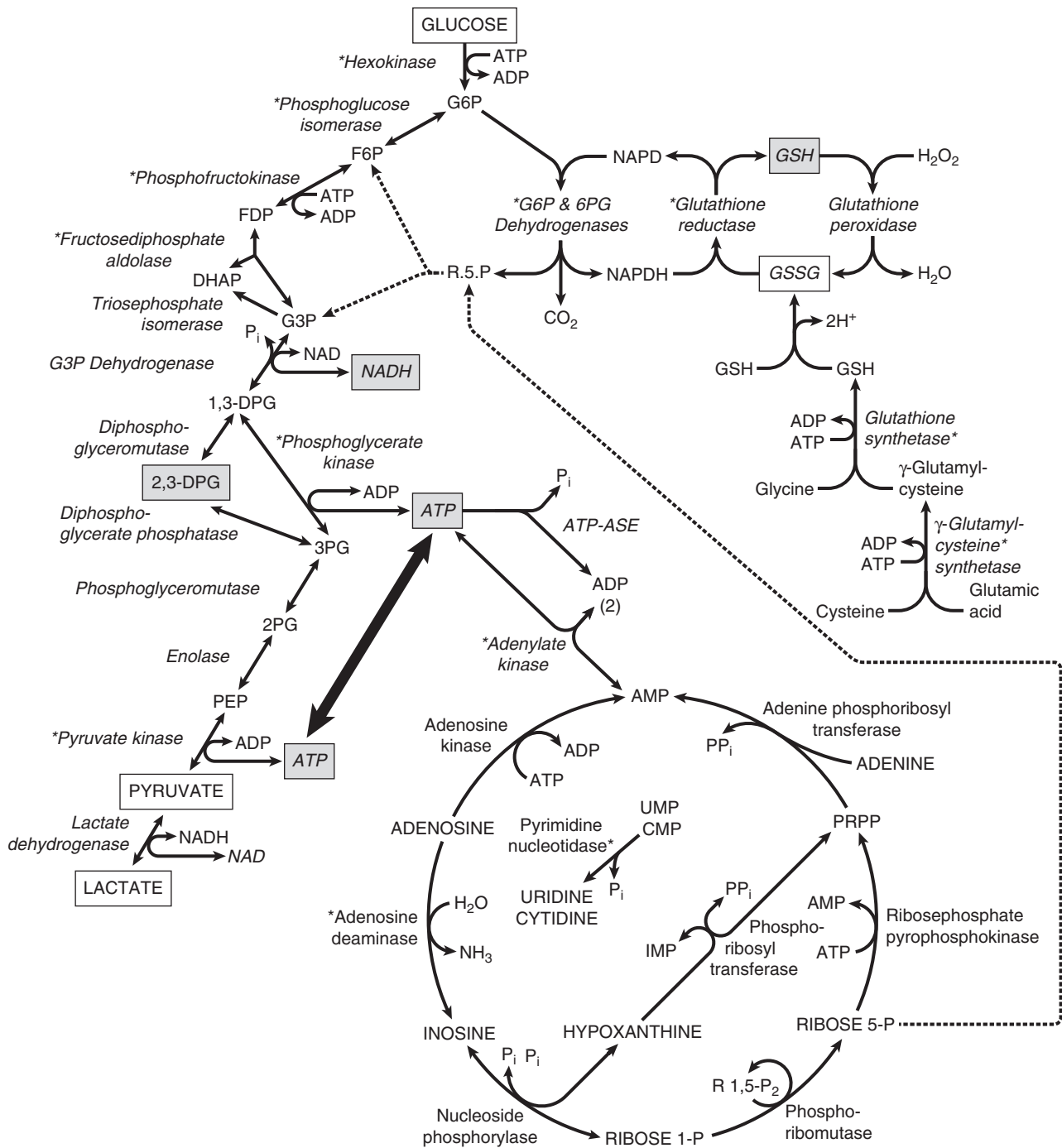


FIGURE 27.2 ■ Pathways of energy metabolism in human erythrocytes. Glucose 6-phosphate (G6P) may be degraded anaerobically to two molecules of lactate via the Embden–Meyerhof (glycolytic) pathway (on the left) or oxidatively via the dehydrogenases of the pentose phosphate pathway. Ribose 5-phosphate (R5P) can re-enter anaerobic glycolysis after conversion to fructose 6-phosphate (F6P) and glyceraldehyde 3-phosphate by enzymes of the terminal pentose phosphate pathway and is also a product of adenosine or inosine degradation. 2,3-Diphosphoglycerate (2,3-DPG) may be generated instead of ATP by diversion of triose through the Rapoport–Luebering shunt. Glutathione may be directly synthesized from constituent amino acids, and its cycling from oxidized (GSSG) to reduced form (GSH) is dependent on generation of reduced pyridine cofactor (NADPH). Asterisks (*) indicate those enzymes found to be defective in association with hereditary haemolytic anaemia. Abbreviations: ADP, adenosine diphosphate; AMP, adenosine monophosphate; ATP, adenosine triphosphate; ATPase, adenosine triphosphatase; CMP, cytidine monophosphate; DHAP, dihydroxyacetone phosphate; 1,3-DPG, 1,3-diphosphoglycerate; 2,3-DPG, 2,3-diphosphoglycerate; FDP, fructose 6-phosphate; G3P, glyceraldehyde 3-phosphate; G6P, glucose 6-phosphate; GSH, reduced glutathione; GSSG, oxidized glutathione; IMP, inosine monophosphate; NAD, nicotinamide adenine dinucleotide (oxidized); NADH, nicotinamide adenine dinucleotide (reduced); NADP, nicotinamide adenine dinucleotide phosphate (oxidized); NADPH, nicotinamide adenine dinucleotide phosphate (reduced); PEP, phosphoenolpyruvate; 2PG, 2-phosphoglycerate; 3PG, 3-phosphoglycerate; P_i, inorganic phosphate; PP_i, inorganic pyrophosphate; PRPP, phosphoribosyl pyrophosphate; R1P, ribose 1-phosphate; R1,5-P₂, ribose 1,5-diphosphate; R5P, ribose 5-phosphate; UMP, uridine monophosphate.

TABLE 27.1 The normal requirements and body stores of the vitamins and minerals essential for erythropoiesis

Vitamin	Dietary sources	Site of absorption	Body stores	Adult daily requirement
Iron	Red meats, fish, poultry, fortified cereals, beans, spinach	Duodenum and proximal jejunum	3–4 g	Male 1 mg Female 1.6 mg
Folate	Dark green leafy vegetables, liver, dried beans and peas, fortified flours and cereals	Duodenum and upper jejunum	10 mg	100 µg
B ₁₂	Liver, kidney, shellfish, and other foodstuffs of animal origin	Distal ileum	2–3 mg	1 µg

the function of ribonucleotide reductase. Many of these roles are aided by the fact that iron may exist in two stable oxidized states, as soluble ferrous (Fe^{2+}) iron and as insoluble ferric (Fe^{3+}) iron.

Iron requirements. The total body content of iron in normal adults is 3–4 g (see Table 27.1). Approximately two-thirds is contained in haemoglobin and the remainder in myoglobin, iron containing enzymes and the reticuloendothelial cells. Approximately 30 mg of iron is required per 24 h for haemoglobin in new red cells and the majority of this is supplied from the reticuloendothelial macrophages that recycle the iron from senescent red cells. However, approximately 1 mg of iron is lost from the body every 24 h and needs to be replaced from dietary sources. Women of childbearing age have additional blood loss from menses or pregnancy, resulting in a further 0.5 mg average loss of iron per 24 h. Infants and adolescents require additional iron during periods of rapid growth. A normal diet provides approximately 15 mg of iron but most is in the form of insoluble iron complexes and only 5–10% is absorbed.

Iron absorption. Absorption of iron occurs in the proximal duodenum. It is aided by the low pH of the hydrochloric acid secreted into the gastric lumen and the presence of reducing agents such as ascorbic acid, which help maintain the iron in the more soluble ferrous form. In contrast, other agents such as tannins, phytates and phosphates bind the iron within the intestinal lumen, inhibiting its absorption.

Although iron has an essential metabolic function, excess is toxic and therefore it is not surprising that iron metabolism is tightly regulated.

Hepcidin plays a critical role in iron metabolism. This 25 amino acid peptide is produced predominantly in the liver, from an 84 amino acid precursor. Its synthesis is stimulated by iron and inflammation (driven by IL-6) and inhibited by iron deficiency, hypoxia (via HIF) and conditions such as ineffective erythropoiesis where the concentration of growth differentiating factor 15 (GDF-15) in the developing erythroblasts is increased. Hepcidin negatively regulates iron homeostasis by binding to ferroportin: the resultant hepcidin-ferroportin complex is phosphorylated, internalized by the cell, ubiquitinated and subsequently degraded in the lysosomes. Ferroportin is a transmembrane protein that functions as a basolateral exporter of iron, playing an essential role in the release of iron from macrophages,

intestinal enterocytes and placental syncytiotrophoblasts, allowing the uptake and transfer of iron by transferrin to sites of need. Thus the stimulation of hepcidin production results in reduced iron absorption and mobilization (see Fig. 27.3). Genetic haemochromatosis (see Chapter 14), where there is significant iron accumulation and deposition in various body tissues resulting in cirrhosis, cardiac failure and diabetes, is a consequence of a defect in the hepcidin pathway.

Iron transport and storage. Transferrin, a single chain polypeptide produced in the liver, acts as the plasma transporter of iron, binding two molecules of ferric iron per molecule. The transferrin-iron complex binds to transferrin receptors, cell surface glycoproteins consisting of two identical 95 kDa subunits linked by a disulphide bridge, localized on cells dependent on iron for function, e.g. for the production of haemoglobin, myoglobin or iron dependent enzymes. After binding, the complex is internalized and degraded, releasing the iron. Ferritin is the primary iron storage protein, found in all tissues but especially the reticuloendothelial system, with small amounts in the plasma. Each molecule of ferritin can store up to 4000 atoms of iron and consists of 24 subunits of two different types, H and L. Variation in the proportion of H and L subunits results in molecular heterogeneity, for example heart and red cell ferritin contains more H subunits than ferritin from liver and spleen, which are rich in L subunits.

The synthesis of all the proteins involved in iron metabolism is finely controlled at the molecular level by having iron response elements (IREs) within their mRNA. In iron replete situations, iron response proteins (IRPs), present in the cytoplasm of many cells, are non-functional by virtue of the binding of an iron-sulphur complex. However, when the cell is iron deplete, the IRPs bind to IREs to stimulate the gene transcription of proteins that enhance iron absorptive capacity.

Causes of iron deficiency anaemia

Iron deficiency is most commonly secondary to blood loss, usually from gastrointestinal or menstrual sources. In the developing world, hookworm infestation resulting in gastrointestinal blood loss is the commonest cause of iron deficiency anaemia; the resultant reduced working capacity has significant socioeconomic impact. Coeliac disease, which is associated with flattening of the duodenal mucosal villi with subsequent malabsorption, accounts for 5% of cases of iron deficiency. Other causes

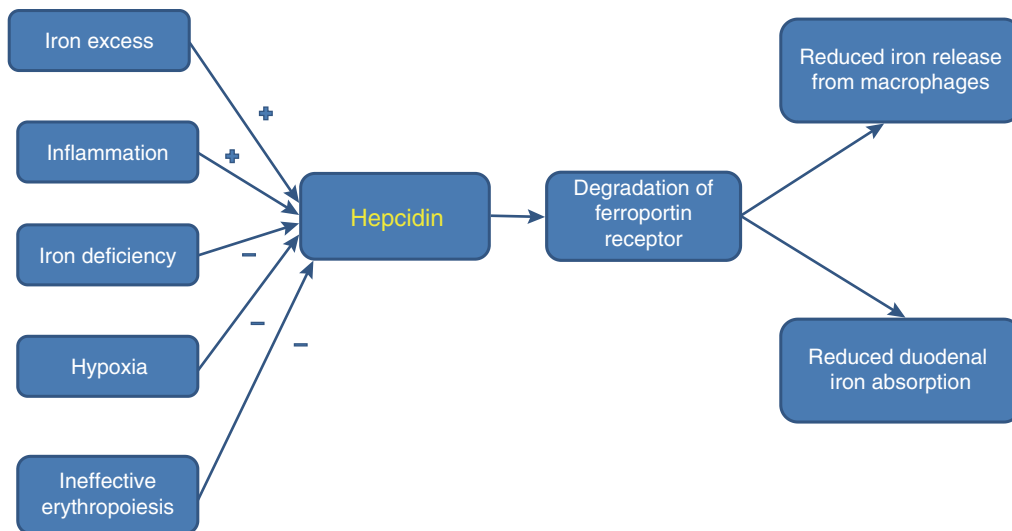


FIGURE 27.3 ■ Factors affecting hepcidin production by the liver and consequent iron absorption and utilization. When hepcidin production is stimulated by conditions such as iron excess or inflammation, it binds to and results in the degradation of ferroportin in enterocytes and macrophages, preventing the absorption and utilization of iron. Conditions associated with iron deficiency or ineffective haemopoiesis inhibit hepcidin production, thereby resulting in iron accumulation.

include inadequate dietary iron during periods of rapid growth such as puberty.

Clinical consequences of iron deficiency

Iron deficiency causes a hypochromic microcytic anaemia (see Chapter 26) with additional clinical features such as koilonychia (a spoon-shaped malformation of the nails), angular stomatitis, glossitis, alopecia and pica. There are concerns that significant iron deficiency in infants results in impaired mental development.

Laboratory determination of iron status

Red cell parameters. Iron deficiency is associated with a reduction in the red cell mean corpuscular volume (MCV) and mean cell haemoglobin concentration (MCHC).

Hypochromic red cells. Some automated analysers allow the determination of numbers of hypochromic red cells. Counts of >6% are suggestive of iron deficiency.

Serum iron. Measurement of serum iron concentration alone provides little useful information of iron status as values show considerable variation within normal individuals. Low concentrations are seen in iron deficiency but are also seen in the anaemia of chronic disease and after surgery.

Serum ferritin. This parameter provides an accurate reflection of iron stores in healthy individuals with normal concentrations ranging from 15 µg/L to 300 µg/L. However, it is an acute phase reactant and its concentration may therefore be falsely normal in iron deficiency with coexistent inflammation.

Serum iron binding capacity, transferrin and transferrin saturation. In the plasma, each molecule of transferrin binds up to two iron ions. Thus, the total iron binding

capacity (TIBC) of plasma reflects serum transferrin concentration. It is raised in iron deficiency and pregnancy. Transferrin saturation is a ratio of serum iron concentration to TIBC expressed as a percentage. A value of <16% is suggestive of iron deficiency whereas elevated fasting transferrin saturations of >55% in males and >50% in females are suggestive of iron overload. In the anaemia of chronic disease, both serum iron and TIBC are reduced, thus the transferrin saturation is usually normal.

Serum transferrin receptor. As erythroblasts mature, cleavage releases the extracellular portion of the transferrin receptor into the plasma. The concentration of the transferrin receptor rises in iron deficiency but not in the anaemia of chronic disease. Transferrin receptor concentration can be measured by enzyme immunoassay, although the assay is only available in a few specialist centres.

Hepcidin. Assays for this protein are under development but as yet are poorly standardized. Controversy exists as to whether urinary or serum hepcidin measurements are superior.

Bone marrow aspiration. At present, staining bone marrow aspirates for iron is the 'gold standard' for assessing iron deficiency.

The megaloblastic anaemias resulting from vitamin B₁₂ and folate deficiency

The megaloblastic anaemias give rise to characteristic morphological appearances: the red cells are macrocytic and hypersegmented neutrophils are present (see Chapter 26). Ineffective erythropoiesis is apparent, manifest in the serum by raised unconjugated bilirubin and reduced haptoglobin concentrations and markedly raised lactate dehydrogenase (LDH) activity. Megaloblastosis arises when either drugs or deficiencies of cobalamin or folate interfere with DNA synthesis.

Folate metabolism

Folates are a group of compounds derived from pteroylglutamic acid (see Fig. 27.4). Excellent dietary sources of folates include green leafy vegetables, liver, yeast and nuts (see Table 27.1).

Folate requirements. A normal diet in the developed world contains about 250 µg of folates per day but <1% of this is in the form of folic acid. The principal naturally occurring folates are tetrahydrofolate (THF), 5-methyl tetrahydrofolate (5-methyl THF) and 10-formyl tetrahydrofolate (10-formyl THF), which are polyglutamated. Total body folate in the adult is 10 mg, providing sufficient stores for four months given a daily metabolic requirement of 100 µg (see Table 27.1).

Absorption of folate. The majority of folate is absorbed in the upper small intestine, where pteroylpolymethylglutamate hydrolase on the mucosal brush border hydrolyses folate polymethylglutamates to the monomethylglutamate form. They are then converted to 5-methyl THF and transported by an active carrier mediated mechanism across the enterocytes and into the portal blood.

Within the plasma, one third of 5-methyl THF is loosely bound to albumin whilst the remainder is unbound. It is taken up by active transport into replicating cells. Here the folates play a vital role by acting as coenzymes in the transfer of single carbon groups in pathways involved in the synthesis of purines, pyrimidines (for DNA and RNA synthesis) and methionine (for methyl donor reactions) (see Table 27.2).

Causes of folate deficiency

- Nutritional.
- Malabsorption: diseases of the small intestine, coeliac disease, tropical sprue.
- Excessive utilization or loss: pregnancy, prematurity, haemolytic anaemias, skin conditions, inflammatory conditions, haemodialysis.
- Antifolate drugs: anticonvulsants, trimethoprim, methotrexate.

Features of folate deficiency

Folate deficiency is thought to cause megaloblastic anaemia by inhibiting thymidylate synthesis, which is required for DNA production. It causes a macrocytic anaemia with hypersegmented neutrophils and a megaloblastic bone marrow. It may be associated with a pancytopenia. For further details on morphology, see Chapter 26.

Laboratory determination of folate status

Both serum and red cell folate concentrations can be measured by immunoassay. Serum folate concentration is affected by recent dietary intake and may be low after only a short period of inadequate diet. Conversely, serum folate concentration may rise in vitamin B₁₂ deficiency owing to a blockage in the conversion of 5-methyl THF, the major circulating form, to THF.

Red cell folate may be a more accurate test of body folate status as its concentration is not affected by recent diet. However, red cell folate concentration is also reduced in vitamin B₁₂ deficiency.

Vitamin B₁₂ metabolism

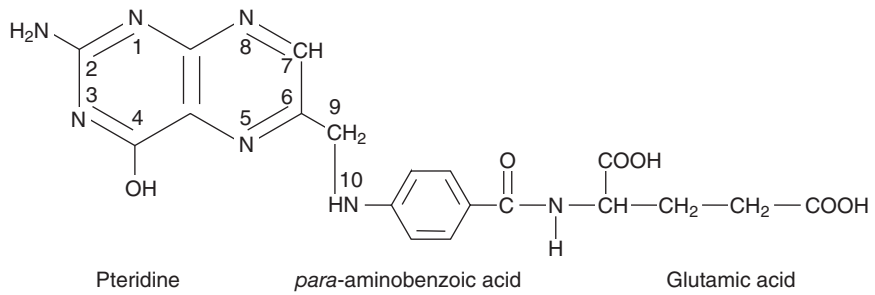
Structure of vitamin B₁₂. The vitamin B₁₂ (cobalamin) molecule is centred on an atom of cobalt; this is the only known function of cobalt in humans. The surrounding corrin ring is made up of four pyrrole units in a similar manner to porphyrins (see Fig. 27.5), with variations in two substitutions above and below the ring giving rise to the various forms of the vitamin. Vitamin B₁₂ is mainly contained within the mitochondria in the 5'-deoxyadenosyl form, where it plays a role in the conversion of L-methylmalonyl-CoA to succinyl-CoA. The other main form, methylcobalamin, is found within the cytoplasm and plasma where it is a cofactor for the conversion of homocysteine to methionine, a vital part of the pathway that creates a universal methyl donor (see Fig. 27.6).

Source of vitamin B₁₂. Vitamin B₁₂ is synthesized only by microorganisms and the only source for humans is food of animal origin. Liver is the richest source of vitamin B₁₂, but it is present in almost all animal products, including milk. No vegetable food source contains significant amounts of vitamin B₁₂ unless contaminated by bacteria (see Table 27.1).

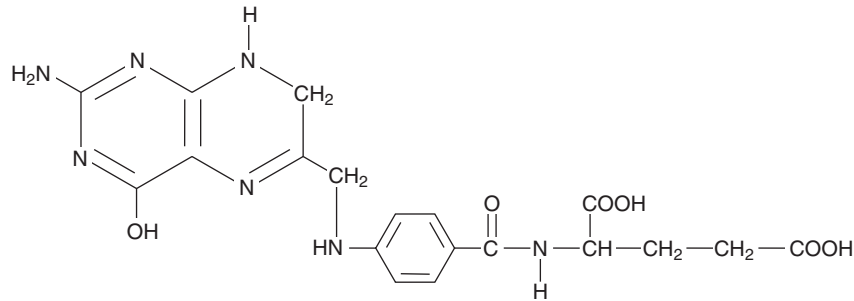
Vitamin B₁₂ requirements. An adult human requires only 1 µg of vitamin B₁₂ per 24 h and has stores of 2–3 mg. Thus, it can take 3–4 years for vitamin B₁₂ deficiency to develop (see Table 27.1).

Absorption of vitamin B₁₂. Dietary vitamin B₁₂ is bound to food proteins and must be freed by gastric acid. The parietal cells of the stomach secrete the glycoprotein intrinsic factor (IF), which combines with vitamin B₁₂ to form a complex, which resists proteolytic digestion. The vitamin B₁₂-IF complex passes through the small intestine until it binds to cubilin, a surface receptor on the enterocytes of the terminal ileum, where it is internalized. The enterocytes of the ileum have a limited capacity to absorb vitamin B₁₂ because of a limited number of receptor sites; about half of a dose of 1 µg of vitamin B₁₂ will be absorbed, the proportion falling markedly with higher doses. The enterocytes have a refractory period of about 6 h before they can absorb any more vitamin B₁₂.

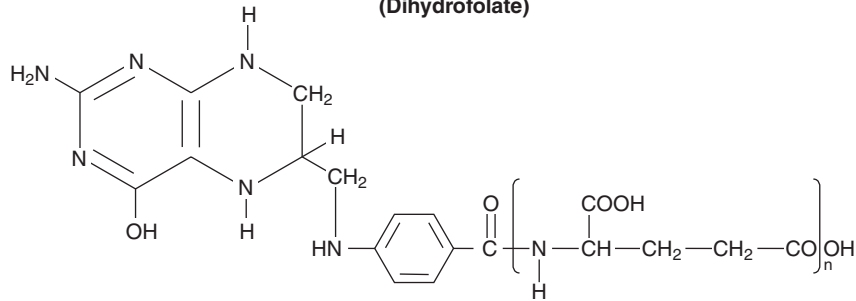
Within the enterocyte, the vitamin B₁₂ is liberated and, after binding to the β-globulin carrier transcobalamin II (TCII), is released into the blood. The TCII-vitamin B₁₂ complex is termed 'holotranscobalamin'. Transcobalamin II readily releases the vitamin B₁₂ to the bone marrow and other tissues, and, for this reason, holotranscobalamin is also described as 'active B₁₂'. Other transcobalamins also exist; transcobalamin I and III derive mainly from specific granules found in neutrophils and bind vitamin B₁₂ tightly, in contrast to TCII, and do not release it into the tissues. Congenital TCII deficiency can occur, the affected infant presenting with megaloblastic anaemia a few weeks after birth.



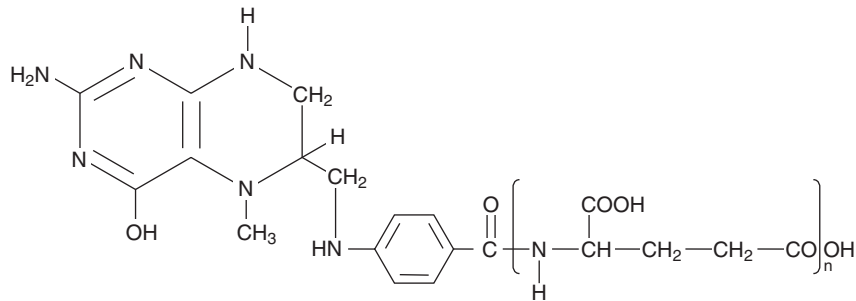
Pteroyl glutamic acid



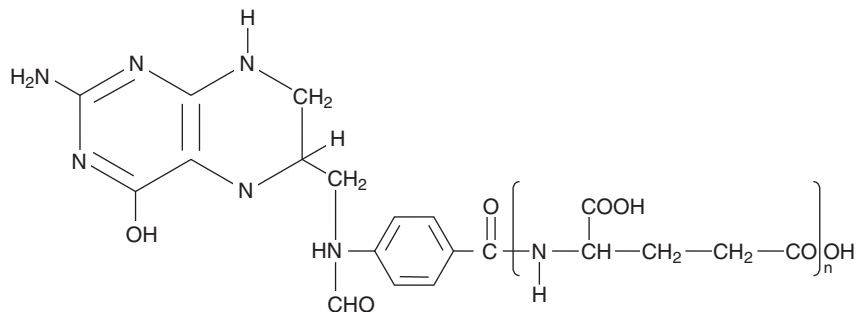
Dihydropteroyl glutamic acid (Dihydrofolate)



Tetrahydropteroyl glutamic acid (Tetrahydrofolate)



5-methyl tetrahydrofolate



10-formyl tetrahydrofolate

FIGURE 27.4 ■ Pteroylglutamic acid and derivatives. The nitrogen atoms in the 5 and 10 positions are the sites involved in single carbon transfer reactions. Tetrahydrofolate and its 5-methyl and 10-formyl derivatives are shown as polyglutamates.

TABLE 27.2 Reactions in which pteroylglutamate acid derivatives are involved

Folate derivative involved as:	Reaction
Single carbon donor	
10-formyl tetrahydrofolate (THF)	Purine synthesis
5,10-methylene THF	Thymine synthesis
5,10-methylene THF	Synthesis of glycine from CO ₂ and NH ₄ ⁺
5-methyl THF	Homocysteine → methionine
Single carbon acceptor	
THF (forms 5,10-methylene THF)	Serine → glycine
THF (forms 5-formimino THF)	Breakdown of histidine

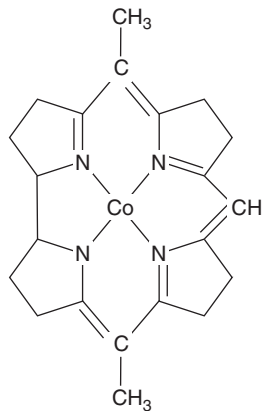


FIGURE 27.5 ■ The corrin ring surrounding an atom of cobalt that forms the core of cobalamin. The other two ligands to cobalt (above and below the plane of the corrin ring), and substituents on the pyrrole units, are not shown.

Causes of vitamin B₁₂ deficiency

- Nutritional: this is rare and is only seen in very strict vegans.
- Malabsorption: the commonest cause of vitamin B₁₂ malabsorption is the autoimmune condition, pernicious anaemia, where gastric atrophy develops secondary to an inflammatory infiltrate. Autoantibodies to gastric parietal cells are seen in 90% of individuals with the condition and 50% develop IF autoantibodies, which either prevent vitamin B₁₂-IF complex formation (binding antibodies) or the subsequent attachment of the vitamin B₁₂ to the enterocyte mucosa (blocking antibodies).
- Gastric causes: such as total or partial gastrectomy.
- Intestinal causes: ileal resection and diseases of the terminal ileum such as Crohn disease or tropical sprue prevent vitamin B₁₂ absorption. Deficiency is also associated with intestinal blind loop syndrome (because of metabolism of the vitamin B₁₂ by the overgrowth of intestinal bacteria) and fish tapeworm (*Diphyllobothrium latum*), which binds cobalamin, preventing its absorption.

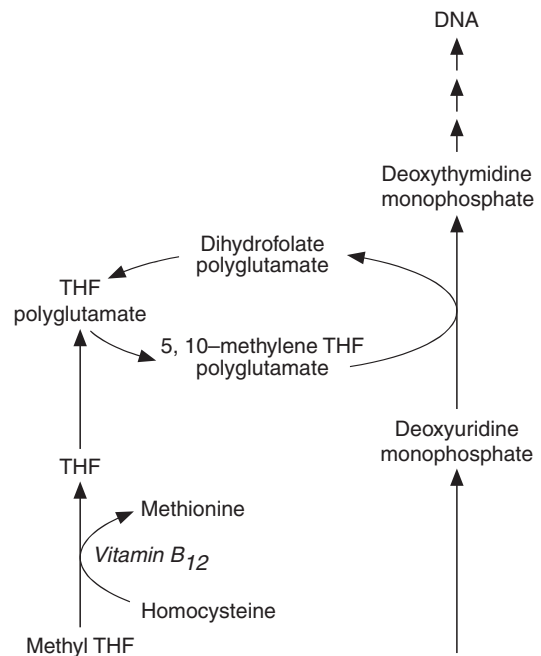


FIGURE 27.6 ■ The link between folate and vitamin B₁₂ deficiency in megaloblastic anaemia. 5,10-Methylene tetrahydrofolate polyglutamate is required for a rate-limiting step in DNA synthesis, the conversion of deoxyuridine monophosphate to deoxythymine monophosphate. Vitamin B₁₂ is required in one of the reactions converting the main circulating form of folate, 5-methyl THF, to 5,10-methylene THF.

- Acquired: prolonged nitrous oxide exposure oxidizes methylcobalamin to an inactive state and results in functional vitamin B₁₂ deficiency. This has been seen in dentists and anaesthetists. Metformin may be associated with low serum vitamin B₁₂ concentrations. The exact mechanism for this is unknown but it has been postulated to be caused by metformin interfering with the calcium dependent channels responsible for the ileal absorption of the vitamin.

Features of vitamin B₁₂ deficiency

Vitamin B₁₂ is a coenzyme in the interconversion of the different forms of folate (see Fig. 27.6), so its deficiency results in a megaloblastic anaemia identical to that seen in folate deficiency. Deficiency may also result in neurological symptoms (e.g. peripheral neuropathy, dysfunction of the posterior columns of the spinal cord and sometimes psychotic illnesses and dementia). The precise mechanism for this has not been elucidated but may be linked with impaired conversion of homocysteine to methionine, resulting in either reduced availability of S-adenosylmethionine impairing sphingomyelin synthesis or in the toxic accumulation of the homocysteine metabolite S-adenosylhomocysteine.

Laboratory determination of vitamin B₁₂ status

Serum vitamin B₁₂. This is usually measured by automated immunoassay. Normal concentrations are 160–1000 ng/L. A low concentration is not specific for vitamin B₁₂ deficiency and may be found in one-third of patients with folate deficiency and in normal pregnancy.

Measurement holotranscobalamin (TCII - bound vitamin B₁₂, or 'active B₁₂') may be a more sensitive and specific indicator of physiologically relevant vitamin B₁₂ deficiency.

Serum methylmalonate and homocysteine. Vitamin B₁₂ deficiency results in the elevation of methylmalonate and homocysteine concentrations (see Fig. 27.6). However, concentrations of both compounds fluctuate and may be raised in renal impairment, smoking and (on single occasions) in up to 30% of normal volunteers, making definition of a specific cut-off level difficult.

Deoxyuridine suppression test. The conversion of deoxyuridine monophosphate (dUMP) to deoxythymidine monophosphate (dTMP) during DNA synthesis is by methyl group transfer, facilitated by both vitamin B₁₂ and folate. The other source of dTMP is via phosphorylation of deoxythymidine, catalysed by thymidine kinase, which is subject to feedback inhibition by its product, dTMP. Normal marrow, pre-incubated with dUMP, successfully converts this to dTMP, and incorporates less subsequently added tritiated thymidine into DNA. If the marrow is deficient in either vitamin B₁₂ or folate, methyl group transfer is reduced, so the cells have greater capacity to incorporate tritiated thymidine into DNA following pre-incubation with dUMP.

Antibody tests. Tests for the presence of antibodies to gastric parietal cells are positive in 90% of patients with pernicious anaemia; however, this test is not specific as it is also positive in around 15% of healthy elderly people. The presence of IF antibodies is more specific but is found in only 50% of patients with pernicious anaemia.

Schilling test. This test, which measured the 24-h urinary recovery of vitamin B₁₂ following an oral radiolabelled dose of the vitamin, is now obsolete.

Anaemias due to reduction in red cell production: inherited causes

Inherited disorders of haemoglobin synthesis, such as sickle cell anaemia, result in ineffective erythropoiesis and a microcytic anaemia. These disorders are covered in Chapter 29.

Other inherited disorders of red cell production are rare and usually present in infancy. Fanconi anaemia is an autosomal recessively inherited disorder of a group of DNA repair proteins. Patients have various somatic deformities, a pancytopenia and markedly increased risk of both solid and haematological malignancies. Diamond-Blackfan syndrome is associated with an anaemia secondary to a reticulocytopenia and carries an increased risk of acute myeloid leukaemia.

Anaemias associated with reduction in red cell production: acquired causes

Anaemia of chronic disease

Chronic diseases are often associated with an elevation of proinflammatory cytokines such as TNF- α and IL-6. These

cytokines are potent inducers of hepcidin, which acts to inhibit intestinal iron absorption and the release of iron from macrophages. Thus iron studies in chronic diseases reveal a low serum iron and iron binding capacity with a raised ferritin. The anaemia of chronic disease is usually normocytic, but in approximately 20% of cases, is microcytic.

Aplastic anaemia

Stem cell failure can be caused by toxins affecting cell production. This may be predictable dose dependent toxicity, such as following exposure to high dose irradiation, cytotoxic drugs and benzene. Rarely, the aplasia occurs as an idiosyncratic reaction to drugs such as non-steroidal anti-inflammatories and chloramphenicol, or as a reaction to viruses such as hepatitis, which is probably immune mediated. It is sometimes associated with an acquired defect in the glycosylphosphatidylinositol (GPI) anchor, which plays an important role in the attachment of a number of cell surface antigens, protecting the cells from complement mediated premature destruction. This rare disorder is called paroxysmal nocturnal haemoglobinuria (PNH) (see p. 529).

Myelodysplasia

The myelodysplastic syndromes are a heterogeneous group of disorders where there is ineffective haemopoiesis, resulting in progressive multilineage cytopenia and an increased propensity to transform to an acute leukaemia. The incidence of these conditions rises with age, with median onset in the seventh decade.

Malignant infiltration of the bone marrow

Infiltration of the marrow causes loss of normal haemopoietic tissue. This gives rise to a characteristic leukoerythroblastic appearance on a blood film. Immature white cells, nucleated red cells and red cells with a teardrop shape appear and the platelet count falls. Leukoerythroblastic anaemia is a characteristic finding when a malignancy has metastasized to the marrow. It also occurs in myelofibrosis, where there is an increase in reticulin fibres in the bone marrow and splenomegaly.

ANAEMIAS ASSOCIATED WITH INCREASED RED CELL LOSS

Bleeding

Acute bleeding, from any source, results in a reticulocytosis and thus, the anaemia may be normocytic or even macrocytic if the reticulocyte count is high, as these cells are larger than mature red cells. With time, iron stores become diminished and a microcytic anaemia develops.

Haemolysis

Haemolysis may be defined as the premature destruction of red cells, that is, the shortening of their normal life span of approximately 120 days. The destruction of the

red cells may be effected by the macrophages in the reticuloendothelial system, for example in the spleen, liver or bone marrow; such destruction is said to be *extravascular*. Alternatively, destruction of red cells may take place within the circulation, i.e. *intravascular* haemolysis. Often there is a combination of both mechanisms. Normally, the bone marrow responds to the anaemia by increasing red cell production, and where this balances the loss of red cells, there is said to be a compensated haemolytic state. If the rate of the destruction exceeds the capacity of the marrow's compensatory increase in erythropoiesis, a haemolytic anaemia ensues.

Thus, haemolytic anaemias are usually characterized by a triad of pathological features:

- increased red cell destruction resulting in *anaemia*
- increased red cell production resulting in *reticulocytosis*
- increased haemoglobin catabolism resulting in *jaundice*.

Laboratory features of haemolysis

In extravascular haemolysis, the haem moiety released from haemoglobin as a consequence of increased red cell destruction is broken down to protoporphyrin, with the liberation of the iron, which enters the plasma pool and is bound to transferrin, increasing its saturation. Protoporphyrin is metabolized to carbon monoxide and biliverdin, which is subsequently further reduced to bilirubin. An increased plasma unconjugated bilirubin concentration to $> \sim 50 \mu\text{mol/L}$ results in jaundice. As unconjugated bilirubin is bound to albumin, it is not excreted in the urine, but there is increased urinary excretion of its metabolic end-product, urobilinogen (see Chapter 13).

In contrast, in intravascular haemolysis, there is release of haemoglobin into the plasma (haemoglobinaemia) and, when the renal threshold for haemoglobin is exceeded, haemoglobin is excreted in the urine (haemoglobinuria). Some of the haemoglobin in the glomerular filtrate is reabsorbed and degraded by the proximal tubular cells, with the iron released being deposited as haemosiderin. As the tubular cells desquamate, the presence of the intracellular haemosiderin iron can be detected in the urinary sediment by Perls reaction, which is positive for up to six weeks after the episode of intravascular haemolysis. The loss of iron may be sufficiently great to result in overt iron deficiency anaemia, in contrast to the iron accumulation seen in extravascular haemolysis.

Free haemoglobin in the plasma readily dissociates from its tetrameric structure to dimeric units consisting of α - and β -subunits. The haemoglobin α chain binds readily to the β chains of haptoglobin, a glycoprotein produced by the liver. The haemoglobin-haptoglobin complex is rapidly cleared by the reticuloendothelial system. In chronic haemolytic states, there is an increased removal of haptoglobin at a rate that exceeds the synthetic capacity of the liver, resulting in low or undetectable plasma haptoglobin concentrations. The return to a normal concentration takes approximately one week following cessation of the haemolytic process.

Any free plasma haem resulting from haemolysis forms complexes with either haemopexin or albumin.

The haem-haemopexin complex is taken up and metabolized by the liver, resulting in a decrease in plasma haemopexin concentration. The haem-albumin complex may be oxidized to methaemalbumin, which has a characteristic absorption spectrum and gives a brownish colour to the patient's plasma. Acute kidney injury is often seen in intravascular haemolysis. Studies have revealed that haem is directly toxic to renal tubular cells and may form intratubular casts with Tamm-Horsfall protein. The mitochondria of renal cells appear particularly susceptible to haem-mediated damage, with resultant impaired mitochondrial oxygen consumption and autophagocytosis. The kidney protects itself during intravascular haemolysis by inducing haemoxygenase-1 and ferritin production, which act to degrade the haem and bind any resultant free iron, respectively (see Fig. 27.7).

Causes of haemolytic anaemias

The disorders that cause haemolysis form a diverse group. Some of them are very common, for example sickle cell anaemia and thalassaemia, which affect millions worldwide, while some are very rare, for example aldolase deficiency, with only a few families reported in the world literature. The clinical features are equally diverse, varying from little or no clinically evident haemolysis as, for example, with hereditary elliptocytosis, to life-threatening acute intravascular haemolysis associated with *Plasmodium falciparum* malaria (blackwater fever). Clinical manifestations other than haemolysis may complicate the situation further; for example, sickle cell anaemia is associated with vaso-occlusive episodes that may result in life-threatening problems such as pulmonary crises, while some of the inherited deficiencies of enzymes such as triose phosphate isomerase and phosphoglycerate kinase are associated with severe neurological disease.

A classification of haemolytic disorders based on the site of haemolysis would identify very few disorders causing only intravascular haemolysis, as the vast majority are associated with predominantly extravascular haemolysis. Therefore, the best classification of haemolytic disorders is based on whether the haemolysis is inherited or acquired, as summarized in Box 27.1.

Inherited haemolytic anaemia

It is convenient to consider inherited red cell defects under three headings: membrane defects, enzymes defects and haemoglobinopathies. The haemoglobinopathies are considered further in Chapter 29.

Membrane defects

The commonest haemolytic anaemia caused by a membrane defect is hereditary spherocytosis. It is usually inherited in an autosomal dominant manner, resulting in haemolysis of varying severity; the majority of patients have a mild anaemia, although severe haemolysis resulting in neonatal kernicterus has been reported. It is characterized by spherocytic, osmotically fragile cells and in most cases is caused by a defect

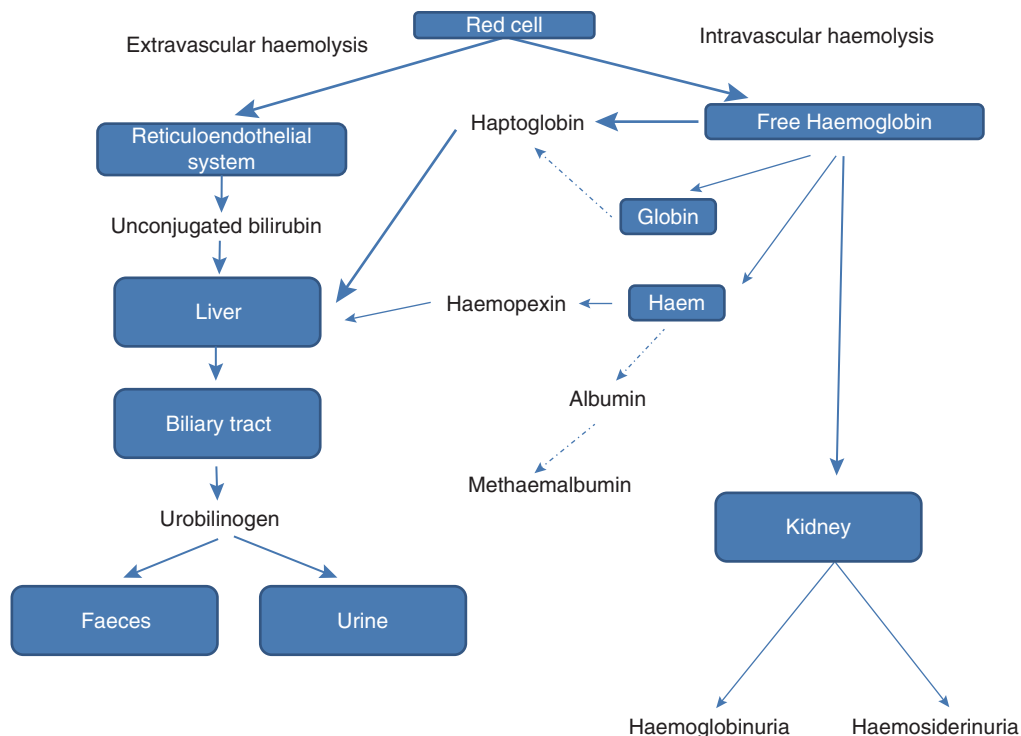


FIGURE 27.7 ■ The pathway of haemoglobin metabolism in haemolytic anaemias. Haemoglobin is liberated from the breakdown of the red cells and becomes bound to haptoglobin. Any free haem binds to haemopexin. Both scavenger proteins are taken up by macrophages within the reticuloendothelial system. When these plasma proteins are consumed, the haem may be either free in the plasma or bound to albumin.

in the ankyrin–spectrin complex, although defects in other membrane proteins can also cause the condition. Extravascular haemolysis occurs predominantly in the spleen because of the poor deformability of the cells. Splenectomy is sometimes considered in patients, as prolongation of the red cell life span results in amelioration of the clinical features.

Hereditary elliptocytosis is a more heterogeneous disorder characterized by large numbers of elliptically shaped red cells in the peripheral blood. It is caused by defects in the spectrin tetramers (see p. 517). There is an increased incidence in West Africa, possibly because of protection from infection by *Plasmodium falciparum*; in vitro studies have demonstrated reduced ability of this organism to parasitize red cells with the defects seen in hereditary elliptocytosis.

Hereditary stomatocytosis (HS) is rare and is associated with the appearance of ‘mouth like’ red cells on blood film. These cells are leaky to cations: hence the condition is associated with pseudohyperkalaemia. Splenectomy should be avoided, owing to the risk of thrombosis in these individuals. The variant termed dehydrated HS, or hereditary xerocytosis, may be associated with perinatal ascites, which resolves spontaneously within the first year of life.

Enzyme defects

These may be grouped under three headings:

- disorders of the pentose phosphate pathway and related enzymes of glutathione metabolism

- disorders of anaerobic glycolysis
- disorders of nucleotide metabolism.

Disorders of the pentose phosphate pathway and related enzymes of glutathione metabolism. The pentose phosphate shunt provides the reducing power of the red cell in the form of NADPH, which maintains glutathione in the reduced form (GSH) via the closely linked glutathione pathway. The GSH protects the red cells from oxidative damage; inadequate supplies result in peroxidation of the red cell membrane, denaturation of haemoglobin and its precipitation as Heinz bodies, resulting in reduced cell deformability and intravascular haemolysis.

The commonest disorder of the pentose phosphate pathway is a defect of the enzyme glucose 6-phosphate dehydrogenase (G6PD), affecting 400 million people worldwide. The gene for the enzyme is located on the X chromosome. The wild type is *G6PD(B)*. More than 300 variants have been described with differences in maximum enzyme activity and/or gene expression. The best known of these affects people of African origin, *G6PD(A-)*, or from Mediterranean countries, *G6PD(Med)*. Most individuals with G6PD deficiency are asymptomatic but liable to have acute haemolytic crises under oxidative stress from certain drugs (see Table 27.3), ingestion of fava beans or acute infection. Favism has been described in Mediterranean countries since classical times.

The cause of the haemolysis is a failure to produce enough GSH to protect the red cell membrane from

BOX 27.1 Classification of haemolytic disorders**Congenital****Red cell membrane defects**

- Hereditary spherocytosis
- Hereditary elliptocytosis
- Acanthocytosis in abetalipoproteinaemia
- Hereditary stomatocytosis
- Rh_{null} disease

Red cell enzyme defects

- Disorders of pentose phosphate pathway and related enzymes of glutathione metabolism
 - Glucose 6-phosphate dehydrogenase
 - Glutamylcysteine synthetase
 - Glutathione synthetase
 - Glutathione reductase
- Disorders of anaerobic glycolysis
 - Pyruvate kinase
 - Hexokinase
 - Phosphoglucose isomerase
 - Phosphofructokinase
 - Triose phosphate isomerase
 - Phosphoglycerate kinase
- Disorders of nucleotide metabolism
 - Pyrimidine 5'-nucleotidase deficiency
 - Adenosine deaminase hyperactivity
 - Adenosine triphosphatase deficiency
 - Adenylate kinase deficiency

Haemoglobinopathies

- Globin chain synthesis, i.e. thalassaemias
- Globin chain structure

Acquired**Immune haemolytic anaemias**

- Incompatible blood transfusion
- Haemolytic disease of the newborn

- Autoimmune (warm reacting antibody)
 - Idiopathic
 - Infection, e.g. *Mycoplasma pneumoniae*
 - Lymphoproliferative disorders
 - Other malignant disorders
 - Immunodeficiency states
 - Systemic lupus erythematosus (SLE) and other autoimmune disorders
 - Drugs
- Autoimmune (cold reacting antibody)
 - Cold haemagglutinin disease (idiopathic or secondary)
 - Paroxysmal cold haemoglobinuria

Non-immune haemolytic anaemia

- Infections
 - Protozoal, e.g. malaria, toxoplasmosis, leishmaniasis
 - Bacterial, e.g. clostridial infection, bartonellosis
- Traumatic and microangiopathic disorders
 - Prosthetic cardiac valves
 - Haemolytic uraemic syndrome
 - Thrombotic thrombocytopenic purpura
 - Disseminated intravascular coagulation
 - Hypertension, including pregnancy-associated hypertension
 - March haemoglobinuria
 - Thermal injury
- Chemical agents
 - Oxidant drugs (see Table 27.3)
 - Non-oxidant agents, e.g. water, copper
 - Venoms
- Disorders of the red cell membrane
 - Vitamin E deficiency in neonates
 - Spur cell anaemia in chronic liver disease
 - Paroxysmal nocturnal haemoglobinuria

TABLE 27.3 Drugs and chemicals associated with significant haemolysis in subjects with G6PD-deficiency

Drugs	Definite association	Possible association	Doubtful association
Antimalarials	Primaquine Pamaquine Pentaquine	Chloroquine	Quinacrine Quinine
Sulfonamides	Sulfanilamide Sulfacetamide Sulfapyridine Sulfamethoxazole	Sulfamethoxypyridazine Sulfadimidine	Sulfoxone Sulfadiazine Sulfamerazine Sulfisoxazole
Sulfones	Thiazolesulfone Diaminodiaphenylsulfone (dapson)		
Nitrofurans	Nitrofurantoin		
Antipyretic analgesics	Acetanilid		Aminopyrine Paracetamol (acetaminophen) Phenacetin Aspirin
Others	Nalidixic acid Naphthalene Niridazole Phenylhydrazine Toluidine blue Trinitrotoluene (TNT) Methylthioninium chloride Phenazopyridine	Chloramphenicol Vitamin K analogues	4-aminosalicylic acid (PAS) L-DOPA Vitamin C Dimercaprol Doxorubicin Probenecid

oxidative stress. Glucose 6-phosphate dehydrogenase is the first and most important enzyme in the pentose phosphate pathway that generates the NADPH that is required for glutathione reduction. Normal *in vivo* flux uses only a small percentage of the maximum enzyme capacity. This explains why G6PD variants with low residual activities are compatible with almost normal function of the red cell. Only under oxidative stress is the protective capacity overwhelmed, leading to predominantly extravascular haemolysis.

In a relatively small number of sporadic cases, the G6PD variant is functionally so abnormal that there is chronic extravascular haemolysis even without any additional oxidative stress.

Disorders of anaerobic glycolysis. Deficiencies of the glycolytic enzymes hexokinase (HK), phosphoglucose isomerase (PGI), phosphofructokinase (PFK), aldolase (ALD), triose phosphate isomerase, phosphoglycerate kinase (PGK), enolase and pyruvate kinase (PK) have all been reported in association with chronic haemolytic anaemia. Almost all cases reported are inherited in an autosomal recessive mode except for PGK deficiency, which is X-linked. All are relatively rare; the commonest in this group is PK deficiency, which has a prevalence of ~1:20000.

Because the glycolytic pathway is mainly concerned with the generation of ATP, it has been proposed that the primary pathogenic defect is low red cell ATP concentration causing rigidity of the red cells. However, the expected reduction in ATP concentration is not a constant finding. This may be partly explained by the presence of increased numbers of reticulocytes, which have higher concentrations of ATP, and/or the preferential destruction of the most metabolically affected red cells, giving a distorted picture of the concentration of the ATP at the time of investigation. It is likely, however, that a number of different factors contribute to the haemolytic process depending on the various metabolic abnormalities resulting from the enzyme deficiency.

Disorders of nucleotide metabolism. The most important abnormalities of nucleotide metabolism that are clearly associated with chronic haemolytic anaemia are deficiency of pyrimidine 5'-nucleotidase (P5N) and hyperactivity of adenosine deaminase (ADA).

The mechanisms of haemolysis in these conditions have not been established, although some interesting hypotheses have been proposed as logical explanations of the biochemical findings. In patients who are homozygous for P5N deficiency, there is an accumulation of pyrimidine nucleotides in the red cells. This could affect the metabolism of the cells by competing with adenine nucleotides, which are the normal cofactors of a number of important enzyme reactions, and by interfering with the function of the pentose phosphate pathway.

In some cases of ADA hyperactivity, there is a low red cell concentration of ATP. It has been hypothesized that this leads to depletion of adenosine through irreversible deamination, so that insufficient adenosine is available for the salvage pathway via adenosine kinase to replenish normal losses from the adenine nucleotide pool.

Acquired haemolytic anaemias

The acquired haemolytic anaemias can be divided into two main groups:

- immune haemolytic anaemias, where the production of antibodies mediates red cell destruction
- non-immune haemolytic anaemias, which are of numerous diverse causes.

Immune haemolytic anaemias

These may be caused by autoantibodies produced against epitopes on the surface of the patient's own normal red cells or by the development of alloantibodies, antibodies produced by the patient against foreign antigens introduced through blood transfusions or neoantigens that develop on red cells during drug induced haemolysis.

The autoimmune haemolytic anaemias are characterized by a positive Coombs test, which detects antibodies, with or without complement, bound to the surface of the red cell. These antibodies are often directed against the rhesus antigen system; they may be of any immunoglobulin subclass and display different thermal activities, hence the term 'warm' and 'cold' acting antibodies. The resulting haemolysis is usually extravascular in nature because of the ability of the autoantibody to fix complement, allowing opsonization and subsequent uptake by the macrophages of the spleen and liver. However, some autoantibodies produced are able to activate the complement pathway through to cell lysis, resulting in a more severe intravascular haemolysis.

Autoimmune haemolytic anaemias may occur without an obvious underlying cause (primary autoimmune haemolytic anaemia) or may be secondary to other conditions such as lymphoma, chronic lymphocytic leukaemia and autoimmune diseases including systemic lupus erythematosus and rheumatoid arthritis.

Non-immune haemolytic anaemias

The main causes of non-immune haemolysis are infections, traumatic and microangiopathic disorders, chemical and physical agents and acquired disorders of the red cell membrane.

Infections. The most important infection associated with haemolytic anaemia is *Plasmodium falciparum*, in which both extravascular haemolysis, from destruction of parasitized erythrocytes in the reticuloendothelial system, and intravascular haemolysis, when the parasites break out of cells, are seen. Rarely, the latter may be extreme enough to produce the dramatic picture known as blackwater fever, with severe intravascular haemolysis associated with a high percentage of parasitized erythrocytes, haemoglobinuria and acute kidney injury.

Other protozoal and bacterial infections that can be associated with haemolysis are listed in [Box 27.1](#).

Traumatic and microangiopathic disorders. The common mechanism in these conditions is contact between red cells and an abnormal surface in the circulation. Cardiac prosthetic valves and intracardiac patch repairs

have been associated with haemolytic anaemia, which is intravascular, though usually mild.

Microangiopathic haemolytic anaemia is a term used to describe a variety of conditions in which there is increased fibrin deposition in the microcirculation. This results in increased red cell fragmentation either directly by damage to red cells by fibrin strands or indirectly by the stress imposed by other cells flowing past cells trapped on fibrin strands. The end-result is intravascular haemolysis. Such conditions include haemolytic uraemic syndrome, which is seen mainly in childhood and in which renal failure is a prominent feature, and thrombotic thrombocytopenic purpura (TTP), which is associated with the diagnostic pentad of various neurological signs, red cell fragmentation, thrombocytopenia, purpura and renal failure. In haemolytic anaemia associated with hypertension (including pregnancy-associated hypertension), there is little fibrin deposition and it is postulated that there is increased fragmentation of red cells trapped at the endothelial surface as a result of the shearing stress of the flow of arterial blood. March haemoglobinuria is a rare and benign condition in which there is red cell destruction in the feet caused by mechanical trauma from walking or running over long distances.

Acquired disorders of the red cell membrane.

Paroxysmal nocturnal haemoglobinuria (PNH) is a rare acquired disorder of the red cell membrane. It is associated with an acquired somatic mutation in the *PIGA* gene on the X chromosome. This encodes the enzyme complex phosphatidylinositol N-acetyl-aminyl-transferase subunit A (PIGA), which plays a pivotal role in the biosynthesis of glycosylphosphatidylinositol (GPI) anchors. These are responsible for the attachment of a large number of surface antigens to the cell membrane, including proteins that protect the cell membrane from complement-mediated attack. Thus GPI anchor deficiency renders the red cells susceptible to complement mediated cell lysis and intravascular haemolysis ensues. The condition is associated with the clinical triad of intravascular haemolysis, bone marrow failure and thrombosis.

Zieve syndrome is an uncommon disorder seen in alcoholics, in which intravascular haemolysis is possibly a consequence of changes in the lipid composition of the cell membrane. Vitamin E deficiency in neonates results in the loss of the protective antioxidant effect of this vitamin and an oxidative haemolysis ensues.

DIAGNOSIS OF HAEMOLYSIS

The search for evidence supporting the presence and cause of haemolysis can be considered in terms of clinical evidence and the results of laboratory investigations.

Clinical evidence

The clinical history should pay particular attention to the onset of jaundice and to any precipitating factors. The colour of the urine may suggest the presence of haemoglobinuria, suggesting intravascular haemolysis. Anaemia

is often asymptomatic, either because of its mild degree or because of its chronicity. However, severe symptomatic anaemia is likely to be of recent onset and suggests an acquired aetiology. The general medical history, including previous surgery, is important but particular attention should be paid to travel, drug history and exposure to chemicals at work or recreation.

A family history may help to establish an inheritance pattern. The majority of families with congenital non-spherocytic haemolytic anaemias caused by enzymopathies show an autosomal recessive transmission, while most non-spherocytic anaemias with unstable haemoglobin or red cell membrane defects show an autosomal dominant pattern. The most important exceptions are the X-linked deficiencies of G6PD and PGK. In these, the affected individuals are hemizygous males or, very infrequently, homozygous females. The vast majority of heterozygous females are clinically unaffected and difficult to detect.

General physical examination provides additional information. In thalassaemia major, frontal bossing and prominence of the maxillae produce a characteristic facial appearance (see p. 553). Splenomegaly is often a prominent finding, particularly in the congenital haemolytic anaemias.

Laboratory investigations

These should be directed to answering two basic questions: is there any haemolysis and if so, what is the mechanism and cause?

Laboratory investigations for the presence of haemolysis

Red cell morphology. In haemolytic anaemias, there is a varying degree of anaemia accompanied by a reticulocytosis. The underlying cause of haemolysis may be suggested by the presence of specific morphological abnormalities (see Chapter 26), for example, spherocytes in both hereditary spherocytosis and autoimmune haemolytic anaemia, and red cell fragmentation in microangiopathic haemolytic anaemia. Heinz bodies, which are composed of denatured haemoglobin, are seen in G6PD deficiency, and some unstable haemoglobins, formed after exposure to oxidant drugs and chemicals, are revealed using the supravital stain crystal violet. Haemoglobin H inclusions (seen in α thalassaemia) can be demonstrated using brilliant cresyl blue. Excessive basophilic stippling, indicating the presence of undegraded RNA, suggests a block in the catabolism of RNA during reticulocyte maturation and is a constant finding in pyrimidine 5'-nucleotidase deficiency.

Total and unconjugated bilirubin. Haem breakdown results in the release of hydrophobic unconjugated bilirubin that is transported, bound to albumin, to the liver where it is conjugated. Therefore, haemolysis results in an increase in plasma unconjugated bilirubin concentration.

Haptoglobin. Any free haemoglobin released from red cell breakdown complexes with haptoglobin and is then cleared by the reticuloendothelial system. A decrease in

the plasma haptoglobin concentration occurs when the daily haemoglobin turnover is doubled, irrespective of whether the haemolysis is extravascular or intravascular. Haptoglobin is more rapidly depleted in intravascular haemolysis. However, congenital ahaptoglobinaemia is seen in 2% of the Caucasian population and concentrations may also be lowered in megaloblastic anaemia (because of ineffective erythropoiesis) and in liver disease. Haptoglobin is an acute phase protein, its concentration rising in many inflammatory conditions, in pregnancy and during use of oral contraceptives and corticosteroids. Under these circumstances, a normal haptoglobin concentration does not exclude haemolysis.

Haemopexin. Haemopexin binds free haem in plasma and in severe intravascular haemolysis, when haptoglobin is depleted, plasma haemopexin falls to very low concentrations or becomes undetectable. However, in mild haemolysis, plasma haptoglobin concentration may be reduced or undetectable but plasma haemopexin concentration normal or only slightly reduced. Low concentrations of haemopexin may be found in renal or liver disease without any evidence of haemolysis and high concentrations are seen in diabetes mellitus, infection or carcinoma. This assay is only available at a few specialized centres.

Methaemalbumin. This is found in plasma when haptoglobin is depleted in the more severe intravascular haemolytic anaemias. It may be detected by the Schumm test: methaemalbumin reduced by the addition of ammonium sulfide shows an intense absorption band in the green spectrum.

Free haemoglobin. Free haemoglobin does not appear in the plasma until all available haptoglobin has been depleted, so it may be undetectable in mild degrees of haemolysis. Therefore, an increased plasma free haemoglobin concentration indicates a significant degree of intravascular haemolysis, provided that the possibility of *in vitro* lysis of red cells during red cell sampling or processing has been excluded. Free haemoglobin in the urine (haemoglobinuria) is also a good indicator of intravascular haemolysis. False positives may occur owing to haematuria (intact red cells in urine) or myoglobinuria, which must be excluded.

Haemosiderinuria. After glomerular filtration, some free haemoglobin is absorbed by renal tubular cells and then broken down. The released iron is deposited as insoluble haemosiderin and is ultimately excreted in the urine when desquamation of the tubular cell occurs. Haemosiderin can be demonstrated within these cells by staining a cytospin preparation of the urine by the Perls technique. Haemosiderinuria can be detected for several weeks after a haemolytic episode, even when there has not been overt haemoglobinuria. Thus, the demonstration of intracellular haemosiderin in the urine is a good sign of mild intravascular haemolysis.

Red cell survival. This can be measured using radiolabelled (usually with ^{51}Cr) autologous red cells. Daily measurements can allow for the calculation of red cell

survival, usually combined with surface counting over the liver and spleen to estimate the relative contribution of each reticuloendothelial organ to the red cell destruction. If uptake of radiolabelled red cells is seen predominantly in the spleen, removal of this organ should ameliorate the haemolysis. This test is performed only in specialist centres.

Laboratory investigations for the cause of haemolysis

After the diagnosis of haemolysis has been made, further investigations are required to establish the precise cause of the reduced red cell survival. These should be directed along a logical pathway to avoid requests for inappropriate tests.

Coombs test (direct antiglobulin test). This is used to detect the presence of red cells coated with immunoglobulin or complement, which is the hallmark of immune haemolysis. The test can be modified to allow the detection of specific immunoglobulin subclasses or complement fractions by the use of specific antisera.

Tests for abnormal haemoglobin. The screening tests for haemoglobinopathy may be described under three main headings, indicating the main haemoglobin property being investigated:

- structural abnormalities, e.g. electrophoresis, isoelectric focusing, high performance liquid chromatography
- functional abnormalities, e.g. sickling test, demonstration of Heinz bodies
- unbalanced haemoglobin synthesis, e.g. quantitation of haemoglobin A₂ and haemoglobin F.

The detailed investigation of haemoglobinopathies is discussed in Chapter 29.

Osmotic fragility tests. These tests are no longer recommended in the routine assessment of haemolytic anaemias, having been replaced by flow cytometry techniques (see below). They are based on spectrophotometric assessment of the lysis of red cells after incubation in water or hypotonic saline for variable periods of time. Normal red cells, as a consequence of their biconcave shape, have an ability to take up more water before they lyse. Spherocytes, because of their high volume to surface area, have a very limited ability to take in water and thus lyse more readily. Conversely, cells with a reduced volume to surface area ratio, such as the cells in thalassaemia or iron deficiency, and reticulocytes, are relatively resistant to lysis. The tests are usually normal in enzyme defects apart from a tail of reduced fragility from the reticulocytes. A normal osmotic fragility does not exclude a diagnosis of hereditary spherocytosis as it may be normal in 10–20% of cases.

The autohaemolysis test. This measures the spontaneous haemolysis of blood incubated at 37 °C for 24 h. It is a useful screening test in cases of suspected haemolytic anaemia if the blood film is not morphologically suggestive of hereditary spherocytosis. If the test

is entirely normal, then an intrinsic red cell defect is unlikely. Correction of the autohaemolysis by the addition of glucose (as an energy source) suggests a membrane defect, as these are associated with increased glucose consumption because of the increased cation leak through the membrane. In enzyme defects, the addition of glucose has no effect as the cells cannot use this energy source.

Flow cytometry. The use of the eosin 5-maleimide (EMA) binding test is now the recommended first-line investigation for the diagnosis of hereditary spherocytosis, where there is no family history or the blood film appearances are atypical. Eosin 5-maleimide binds to band 3 protein (so named on the basis of its electrophoretic mobility). This protein interacts with both ankyrin and protein 4.2 and, in turn, with the spectrin cytoskeleton, which is disrupted in hereditary spherocytosis, resulting in a decreased flow cytometric fluorescence signal. The test has a high level of specificity (99.1%) and sensitivity (92.7%) and can be performed rapidly.

Flow cytometry has also become the gold standard for the diagnosis of PNH, replacing the acid haemolysis (Ham) test. It uses fluorescent-labelled monoclonal antibodies to detect various GPI-anchored surface antigens, principally CD55 (decay accelerating factor) and CD59 (membrane inhibitor of reactive lysis), whose absence on the surface of red cells is compatible with a diagnosis of PNH.

Tests for enzyme deficiencies

It is useful for laboratories to be able to screen for the common red cell enzyme deficiencies, such as G6PD and PK, and to indicate where the defect lies in the less common disorders, but further detailed investigations are best performed in specialist laboratories. It is important that leukocytes and platelets are removed from the samples as they generally have higher enzyme concentrations than the red cells. Also, it should be remembered that reticulocytes have increased concentrations of many 'age-dependent' enzymes, particularly HK, PK, ALD and pyrimidine 5'-nucleotidase. Therefore, if possible all samples should be tested either alongside a control sample with a similar reticulocyte count or the activity of the enzyme under investigation compared with that of a second 'age-dependent' enzyme.

Glucose 6-phosphate dehydrogenase. Screening tests for this enzyme deficiency depend on the inability of the cells to convert an oxidized substrate, such as NADP, to a reduced form that fluoresces under long wavelength UV light. Red cells with <20% of normal G6PD activity do not fluoresce. Problems in interpreting this test are common, particularly in heterozygous women and affected males with G6PD(A-) deficiency and reticulocytosis. In the latter situation, the presence of increased numbers of reticulocytes, which contain increased amounts of G6PD(A-), may give false negative results. Thus negative results do not always exclude enzyme deficiency, while positive results should be confirmed by quantitative assay.

Pyrimidine 5'-nucleotidase. Screening for this deficiency is based on the different spectrophotometric properties of cytidine nucleotides (pyrimidine nucleotides), which absorb maximally in acidic solutions at 280 nm, whilst adenine and guanine (purine nucleotides) and uridine absorb at 260 nm. Normally, >96% of nucleotides in red cells are purine analogues, but this falls to <50% in cases of pyrimidine 5'-nucleotidase deficiency, where the pyrimidine nucleotides accumulate, causing the absorbance ratio to fall.

Red cell metabolites. 2,3-Diphosphoglycerate (2,3-DPG), ATP and GSH are present in red cells in sufficiently high concentrations (millimolar) to be readily measurable by spectrophotometric techniques in most laboratories, providing valuable information on red cell enzymopathies. An increase in 2,3-DPG occurs in most anaemias. Furthermore, hypoxaemia, alkalosis and hyperphosphataemia all lead to increases in 2,3-DPG, irrespective of anaemia. A low concentration of 2,3-DPG is a useful indicator of an enzymopathy early in the glycolytic pathway. Hexokinase, PFK and PGI deficiencies have been frequently, but not invariably, associated with low concentrations of this metabolite. Extremely low 2,3-DPG concentrations have also been reported in a case of complete diphosphoglyceromutase deficiency, in which there is a high haemoglobin concentration because of an increased oxygen affinity in the absence of haemolysis.

A moderate reduction in GSH concentration has been reported in several patients with congenital nonspherocytic haemolytic anaemia including G6PD deficiency. However, very low concentrations of GSH could also be an indication of deficiency in either of the two enzymes of GSH biosynthesis: glutathione synthetase and γ -glutamylcysteine synthetase. In both enzymopathies, some patients have neurological disease in addition to the haemolytic anaemia, which is exacerbated by oxidative stress. For unknown reasons, high concentrations of GSH are present as an epiphenomenon in pyrimidine 5'-nucleotidase deficiency and several syndromes of dyserythropoietic anaemia and ineffective erythropoiesis. Neonates have higher red cell GSH concentrations than adults.

Glycolytic intermediates. The *in vivo* concentrations of glycolytic intermediates are probably the best available measure of the functional status of the glycolytic enzymes. Theoretically, when a given enzyme is functionally defective, it should cause a metabolic block leading to an accumulation of its precursor metabolites and a relative reduction in the concentration of those after it. However, this is not always seen for various reasons, including the increased activity of the enzymes in reticulocytes, but measurement of the glycolytic metabolites may still provide a useful screening test for a metabolic block in some circumstances. For example, PK deficiency may be suspected because of relatively high concentrations of metabolites preceding the PK step, i.e. phosphoenolpyruvate, 2-phosphoglycerate, 3-phosphoglycerate and 2,3-DPG.

CONCLUSION

In the accurate assessment of a patient presenting with anaemia, the clinical history plays a vital role in distinguishing between congenital and acquired causes. This is further supplemented by utilizing the following investigations:

- examination of blood film morphology
- serum ferritin, vitamin B₁₂ and folate concentrations
- coombs test and reticulocyte count
- haemoglobin electrophoresis.

This strategy should result in a specific diagnosis for the majority of individuals, with only a minority (<5% in the UK) requiring more specialized investigations.

ACKNOWLEDGEMENTS

The author is deeply indebted to Dr Pamela A. Gover, who prepared the chapter on Biochemical aspects of anaemia for earlier editions of this book.

Further reading

Hoffbrand AV, Moss P, editors. *Essential haematology*. 6th ed. Oxford: Wiley-Blackwell; 2011.

A simple overview of clinical haematology for students.

Hoffbrand AV, Catovsky D, Tuddenham EGD et al. *Postgraduate haematology*. 6th ed Oxford: Wiley-Blackwell; 2011.

The main UK reference text for senior clinicians working in haematology.

Bain BJ, Bates I, Laffan MA et al. editors. *Dacie and Lewis practical haematology*. 11th ed. Edinburgh: Churchill Livingstone; 2012.

A practical guide to laboratory procedures and techniques.

The porphyrias: inherited disorders of haem synthesis

Michael N. Badminton • George H. Elder

CHAPTER OUTLINE

INTRODUCTION AND OVERVIEW 533

Biochemistry of haem synthesis 533

Overview of the porphyrias 534

Molecular genetics of the porphyrias 534

PORPHYRIAS PRESENTING WITH ACUTE ATTACKS 538

The autosomal dominant acute porphyrias 538

Rare forms of acute porphyria 542

THE CUTANEOUS PORPHYRIAS 542

Bullous porphyrias 542

Erythropoietic protoporphyria and X-linked dominant protoporphyria 546

SECONDARY DISORDERS OF PORPHYRIN METABOLISM 547

CONCLUSION 548

INTRODUCTION AND OVERVIEW

The porphyrias are a group of eight metabolic disorders that result from inherited or acquired functional abnormalities of enzymes of the haem biosynthetic pathway (Table 28.1 and see Fig. 28.1, below). No disease has yet been associated with defects in the 5-aminolaevulinic acid (ALA) synthase-1 (*ALAS1*) gene that encodes the ubiquitous isoform of ALAS, the first enzyme of the pathway. However, gain of function mutations in the erythroid-specific ALAS (*ALAS2*) gene cause X-linked dominant protoporphyria (XLDPP), whereas loss of function causes X-linked sideroblastic anaemia.

The principal clinical features of the porphyrias are neurovisceral or cutaneous, or both. This chapter approaches the subject from a practical perspective, describing how and why patients present, how they are diagnosed and how patients and their families should be managed. A list of frequently used abbreviations is provided in Box 28.1.

Biochemistry of haem synthesis

Haem is essential for life and is synthesized in all cells, although the major sources are bone marrow (80%) and liver (15%). Haemoproteins include haemoglobin and myoglobin, which are the most abundant; the mitochondrial respiratory cytochromes; enzymes such as catalase and tryptophan pyrrolase, and the cytochrome P450 enzymes that are components of many essential metabolic processes, including the metabolism of xenobiotics.

The haem synthetic pathway comprises eight steps, each catalysed by a specific enzyme, of which the first and last three are mitochondrial and the remainder cytosolic (Fig. 28.1). The first enzyme, ALAS,

catalyses the condensation of succinyl-CoA and glycine into 5-aminolaevulinic acid and is the rate-controlling step in all cells. Regulation of ALAS1 in the liver and other non-erythroid tissues is via inhibition by haem, the end product of the pathway, a characteristic that underlies both the pathogenesis and treatment of acute attacks of porphyria. In erythroid cells, regulation of haem synthesis is iron dependent. The second step is the synthesis of porphobilinogen (PBG), a watersoluble, colourless monopyrrole, catalysed by ALA dehydratase. The third step is the polymerization of four molecules of PBG to form the colourless linear tetrapyrrole, 1-hydroxymethylbilane (HMB), by the enzyme hydroxymethylbilane synthase (HMBS, also known as PBG deaminase). This linear molecule is cyclized into the tetrapyrrole ring structure, uroporphyrinogen III, by the enzyme uroporphyrinogen III synthase.

The porphyrinogens are colourless, non-fluorescent, unstable compounds, which rapidly oxidize to their red-purple porphyrin equivalents. Porphyrins absorb light, are fluorescent and therefore photosensitizing. These properties also make them relatively straightforward to measure in the clinical biochemistry laboratory. Uroporphyrinogen III, which is hydrophilic by virtue of its eight carboxy groups, is converted to coporphyrinogen III by uroporphyrinogen decarboxylase, which catalyses the sequential removal of four carboxyl residues. Two further carboxyl groups are removed in an oxygen-dependent dehydrogenation-decarboxylation reaction catalysed by the enzyme coproporphyrinogen oxidase (CPOX) to form protoporphyrinogen, which is then oxidized to protoporphyrin IX by protoporphyrinogen oxidase (PPOX). Progressive decarboxylation makes these precursors and the corresponding porphyrins increasingly hydrophobic, which determines their

TABLE 28.1 Overview of the porphyrias indicating inheritance, prevalence and main clinical presentation

Disorder	Enzyme	Inheritance	Prevalence (overt disease)	Clinical presentation
Acute porphyrias				
ALA dehydratase deficiency porphyria (ADP)	ALA dehydratase (ALAD)	AR	Unknown	Acute neurovisceral attacks
Acute intermittent porphyria (AIP)	Hydroxymethyl-bilane synthase (HMBS) ^{a,b}	AD	1–2:100 000	Acute neurovisceral attacks
Hereditary coproporphyrria (HCP)	Coproporphyrinogen oxidase (CPOX) ^b	AD	<1:250 000	Bullous photosensitivity/acute neurovisceral attacks
Variagate porphyria (VP)	Protoporphyrinogen oxidase (PPOX) ^b	AD	1:250 000	Bullous photosensitivity/acute neurovisceral attacks
Non-acute porphyrias				
Congenital erythropoietic porphyria (CEP)	Uroporphyrinogen synthase (UROS)	AR	<1:10 ⁶	Bullous photosensitivity
Porphyria cutanea tarda (PCT)	Uroporphyrinogen decarboxylase (UROD)	AD ^c	1:25 000 (20% familial)	Bullous photosensitivity
Erythropoietic protoporphyria (EPP)	Ferrochelatase (FECH)	AR	1:100 000	Acute photosensitivity
X-linked dominant protoporphyria (XLDPP)	ALA synthase-2	X linked	Unknown	Acute photosensitivity

AR, autosomal recessive; AD, autosomal dominant.

^aAlso known as PBG deaminase.

^bEnzyme activities are half normal.

^cIn about 20% of patients.

routes of excretion (see Fig. 28.2). The final step in the pathway is insertion of ferrous iron (Fe^{2+}) into protoporphyrin to form haem, catalysed by ferrochelatase (FECH). In the absence of iron, other divalent cations, such as zinc, may be inserted. Although only the III isomer can progress through the pathway to form protoporphyrin IX and haem, HMB may spontaneously cyclize into the uroporphyrinogen I isomer. This forms a substrate for uroporphyrinogen decarboxylase (UROD) and may be converted to coproporphyrinogen I, but is not metabolized further. Adult reference ranges for porphyrins and their precursors are shown in Table 28.2.

Overview of the porphyrias

Clinically manifest porphyria is always associated with detectable overproduction of haem precursors. Each enzyme deficiency, and the increase in activity in XLDPP, gives rise to a specific pattern of overproduction, which defines the corresponding disease (Table 28.3). Three clinical manifestations may occur: acute neurovisceral attacks, skin lesions or both. Acute attacks of porphyria are always accompanied by overproduction of ALA and, in all but ALA dehydratase deficiency porphyria (ADP), of PBG. Porphyrias causing skin lesions are characterized by overproduction of porphyrins. Four of the eight porphyrias can present with acute neurovisceral attacks: the very rare autosomal recessive ADP and the three autosomal dominant acute porphyrias, acute intermittent porphyria (AIP), hereditary coproporphyrria (HCP) and variagate porphyria (VP). Hereditary coproporphyrria and VP can present with skin photosensitivity or acute attacks, or

both. In the other four porphyrias, two types of photosensitization may occur: accumulation of hydrophobic free protoporphyrin in erythropoietic protoporphyria (EPP) and XLDPP is associated with acute photosensitivity, while accumulation of the more water-soluble porphyrins in porphyria cutanea tarda (PCT) and congenital erythropoietic porphyria (CEP) leads to fragile skin and bullae.

Molecular genetics of the porphyrias

All the porphyrias, apart from the sporadic form of PCT, are single gene disorders that are inherited in autosomal dominant, recessive or X-linked patterns (see Table 28.1). Characteristics and chromosomal locations of individual genes are shown in Table 28.4. The *HMBS* and *UROS* genes are alternatively spliced to produce erythroid and ubiquitous isoforms. Disease-specific mutations that abolish or markedly decrease enzyme activity have now been identified in the genes for all the autosomal dominant porphyrias (Human Gene Mutation Database: www.hgmd.org). In most countries, mutational analysis has revealed extensive allelic heterogeneity, with large numbers of mutations identified in each gene, most of which are present in only one or a few families. The main exceptions are the W198X *HMBS* mutation in Sweden and the R59W *PPOX* mutation in South Africa. In both countries, these mutations have been multiplied by founder effects and account for the high prevalence of AIP in Sweden and of VP among individuals of Afrikaans ancestry in South Africa. The proportion of different types of mutation varies little between the diseases, with missense, nonsense, splice site and frameshift mutations contributing to the overall heterogeneity. Large deletions appear

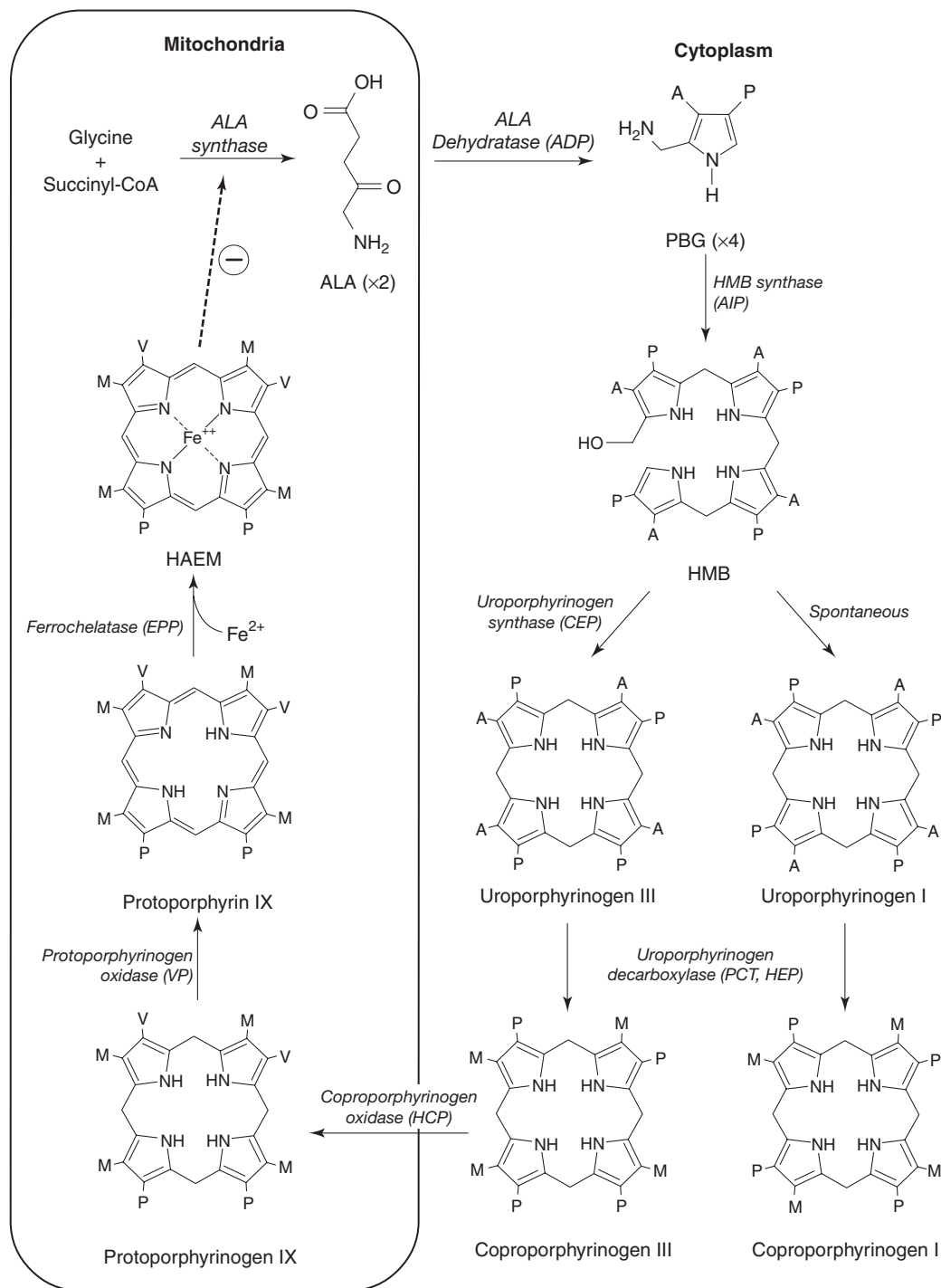


FIGURE 28.1 ■ The haem biosynthetic pathway. The side chains are denoted by: A, acetic acid; M, methyl; P, propionic acid; V, vinyl. For other abbreviations, see [Box 28.1](#).

BOX 28.1 Abbreviations used for porphyrias, and metabolites and enzymes of the haem pathway

ADP	ALA dehydratase deficiency porphyria	HEP	Hepatoerythropoietic porphyria
AIP	Acute intermittent porphyria	HMB	Hydroxymethylbilane
ALA	5-Aminolaevulinic acid	HMBS	Hydroxymethylbilane synthase
ALAD	ALA dehydratase	PBG	Porphobilinogen
ALAS	ALA synthase	PBG-D	Porphobilinogen deaminase
CEP	Congenital erythropoietic porphyria	PCT	Porphyria cutanea tarda
CPOX	Coproporphyrinogen oxidase	PPOX	Protoporphyrinogen oxidase
EPP	Erythropoietic protoporphyria	UROD	Uroporphyrinogen decarboxylase
FECH	Ferrochelatase	UROS	Uroporphyrinogen synthase
HCP	Hereditary coproporphyria	VP	Variegate porphyria
		XLDPP	X-linked dominant protoporphyria

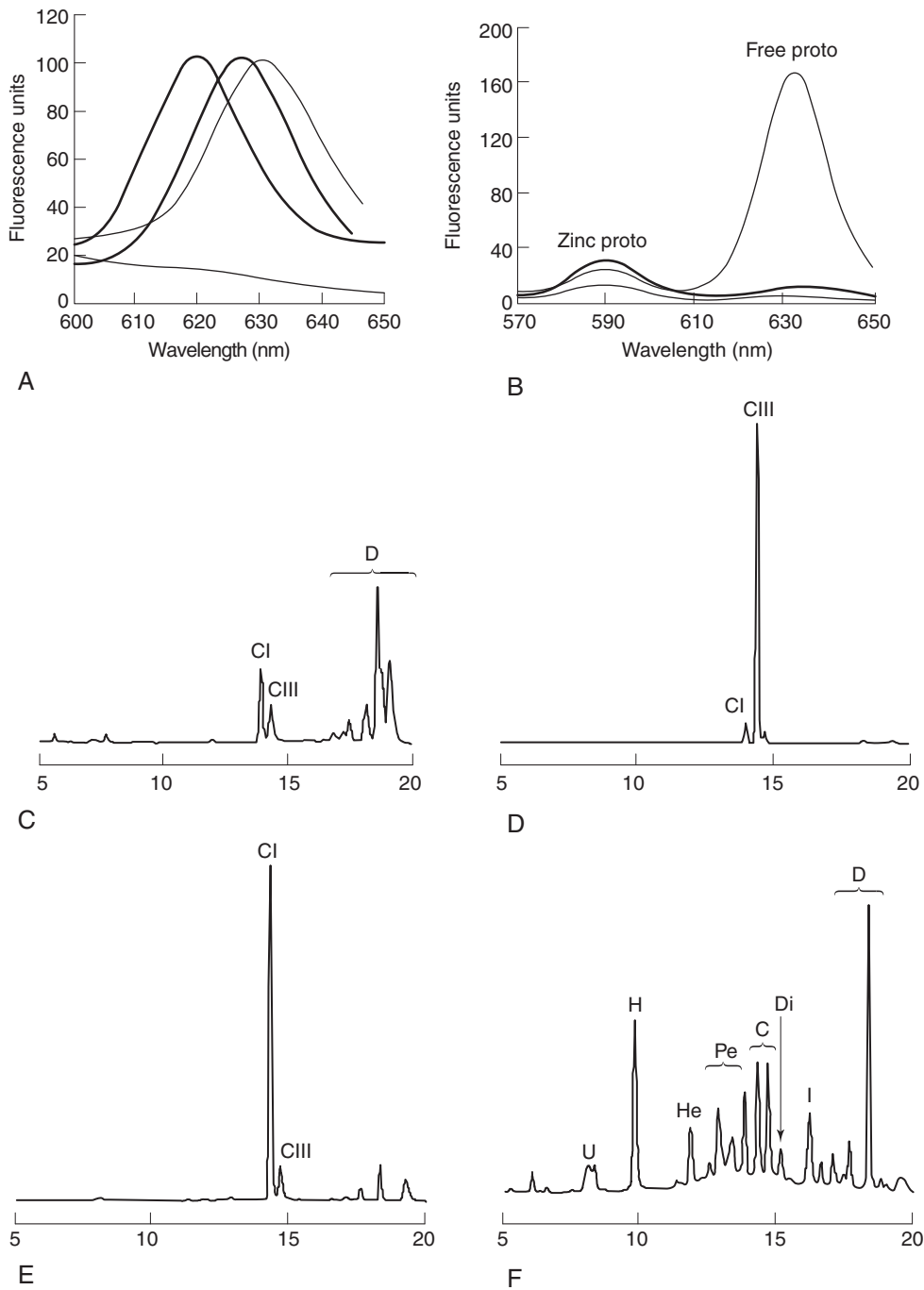


FIGURE 28.2 ■ Examples of key laboratory findings that allow the porphyrias to be distinguished biochemically. (A) Plasma porphyrin fluorescence scans showing the three distinct emission peaks at 620 nm (AIP, CEP, PCT, HCP), 626 nm (VP) and 631 nm (EPP). A negative plasma scan (lower line) is also shown. (B) Fluorescence emission scan of whole blood allows zinc chelated protoporphyrin (proto) (increased in anaemia and lead poisoning) to be distinguished from free protoporphyrin, which is increased in EPP. Markedly increased zinc chelated and free protoporphyrins are characteristic for XLDPP. The lower line indicates a normal scan. High performance liquid chromatography (HPLC) analysis of faecal porphyrins allows three porphyrias with a plasma fluorescence emission peak at 620 nm to be distinguished (faecal porphyrin excretion is normal in AIP). (C) Normal faecal porphyrin HPLC trace indicating predominantly coproporphyrin (C) isomer and dicarboxylic (D) porphyrins (protoporphyrin, pemptoporphyrin, deuteroporphyrin). (D) Faecal porphyrin pattern found in HCP indicating increased coproporphyrin, which is almost entirely the III isomer. (E) Faecal porphyrin trace seen in CEP indicating increased coproporphyrin isomer I. (F) Faecal porphyrin pattern seen in PCT patients indicating the increased excretion of partially decarboxylated intermediates; heptacarboxylic (H), hexacarboxylic (He) and pentacarboxylic (Pe) porphyrins as well as isocoproprophyrin (I) and dehydroisocoproprophyrin (Di), which are pathognomonic for PCT. U, uroporphyrin.

TABLE 28.2 Adult reference ranges for porphyrins and their precursors

Specimen	Metabolite	Reference range
Urine	5-Aminolaevulinic acid	<3.8 µmol/mmol creatinine
	Porphobilinogen	<1.5 µmol/mmol creatinine
	Total porphyrin	<40 nmol/mmol creatinine
	Uroporphyrin	<40%
	Coproporphyrin I	5–40%
Faeces	Coproporphyrin III	40–80%
	Total porphyrin	10–200 nmol/g dry weight
	Coproporphyrin I	5–43%
	Coproporphyrin III	5–37%
	Copro I/III ratio	0.3–1.4
Erythrocytes	Dicarboxylic porphyrin ^a	10–85%
	Total porphyrin	0.4–1.7 µmol/L erythrocytes

^aMainly protoporphyrin but includes pempto-, deuterio- and mesoporphyrins, which are the products of bacterial action on protoporphyrin.

to be uncommon. Perhaps not surprisingly in view of the large number of different mutations, no clinically useful genotype–phenotype correlation has yet been identified in any autosomal dominant porphyria. With current methods of analysis, the sensitivity of mutation identification is over 95% in AIP and VP.

Incomplete penetrance is an important feature of all the autosomal dominant porphyrias; only a proportion of those who inherit a disease-specific mutation develop the disease. In this chapter, the terms latent or presymptomatic will be used to describe individuals who have inherited a porphyria gene but at the time of investigation have not had symptoms. Estimates of penetrance for the autosomal dominant acute porphyrias derived from family studies range from 10% to 40% and are influenced by age and rigour of phenotype definition. In the UK, about 80% of adults identified by family studies as carrying a gene for an acute porphyria never have an acute attack severe enough to require hospitalization. This figure is consistent with the observation that about 30% of patients with AIP present without a family history of the disease, yet family studies almost always reveal

TABLE 28.3 The main biochemical findings in symptomatic porphyria

Porphyria	Urine	Faecal porphyrins	Plasma porphyrins ^a	Erythrocyte porphyrins
Acute porphyrias				
ADP	ALA, coproporphyrin	Normal	Normal	Protoporphyrin (zinc)
AIP	ALA, PBG, uroporphyrin	Normal or slight increase in normal porphyrin pattern	615–620 nm	Normal
HCP	ALA, PBG, coproporphyrin III	Coproporphyrin III (isomer III:I ratio >1.4)	615–620 nm	Normal
VP	ALA, PBG, coproporphyrin III	Protoporphyrin > coproporphyrin III	624–627 nm	Normal
Non-acute porphyrias				
CEP	Uroporphyrin I, coproporphyrin I	Coproporphyrin I	615–620 nm	Protoporphyrin (zinc and free), uroporphyrin I
PCT	Uroporphyrin (I and III), heptacarboxylic porphyrin	Isocoproporphyrin, heptacarboxylic	615–620 nm	Normal
EPP	Normal	Increased protoporphyrin in 40%	626–634 nm	Protoporphyrin (free)
XLDPP	Normal	Protoporphyrin may be increased	626–634 nm	Protoporphyrin (zinc and free)

Porphyria analyses on random urine, faeces and EDTA-preserved blood allow the individual porphyrias to be distinguished. Individual urine and faecal porphyrins are separated and measured by high-performance liquid chromatography with fluorescence detection. Note that EPP cannot be diagnosed by urine porphyrin analysis alone.

^aEmission maximum of plasma fluorescence peak (for examples, see Fig. 28.2).

TABLE 28.4 The molecular genetics of the porphyrias

Disorder	Gene	Chromosome	Gene size (kb)	Exons	cDNA (codons)
ADP	<i>ALAD</i>	9q34	13	13	330
XLDPP	<i>ALAS2</i>	Xp11.21	35	11	550
AIP	<i>HMBS</i>	11q24.1–24.2	10	15	U: 361 E: 344
CEP	<i>UROS</i>	10q25.2–26.3	34	10	U: 265 E: 265
PCT	<i>UROD</i>	1p34	3	10	367
HCP	<i>CPOX</i>	3q12	14	7	354
VP	<i>PPOX</i>	1q21–23	5	13	477
EPP	<i>FECH</i>	18q21.3	45	11	423

U, ubiquitous; E, erythroid.

latent porphyria in their families, de novo mutation being uncommon. Studies of blood donors suggest that the gene for AIP may be present in about 0.06% of the general population. The low clinical penetrance of the acute porphyrias probably reflects a combination of environmental (see below) and genetic influences. The latter have not been identified but are likely to be at loci distant from the disease gene. Factors that determine clinical penetrance in familial PCT are described on page 544.

Clinical penetrance of the autosomal recessive porphyrias, ADP, CEP and EPP is close to 100%, although age at first presentation may be variable. They also show extensive allelic heterogeneity, with most patients whose parents are not consanguineous being compound heterozygotes. Other features of their genetics, and those of XLDPP, are discussed later in this chapter.

PORPHYRIAS PRESENTING WITH ACUTE ATTACKS

The autosomal dominant acute porphyrias

AIP, HCP and VP are autosomal dominant conditions that put the patient at risk of acute neurovisceral attacks, which may be life-threatening. In VP and HCP, skin lesions, indistinguishable from those of PCT, may occur either independently or in combination with acute attacks. In the UK, acute attacks of porphyria are estimated to affect about nine per million of the population, most of whom have AIP. Symptomatic VP is approximately half as common as AIP. HCP is the rarest of these acute porphyrias.

Pathophysiology of acute attacks

Acute neurovisceral attacks occur when hepatic haem synthesis is induced in the presence of deficient HMBS activity, resulting in an accumulation of the haem precursors ALA and PBG. In AIP, this is a primary deficiency, but in VP and HCP, it has been proposed that the deficiency is secondary to inhibition of HMBS by other metabolites in the haem pathway. The exact cause of neuronal damage has not been fully established, but current evidence from liver transplantation suggests that it is due to a hepatic neurotoxin, probably ALA. The neurological lesions comprise axonal degeneration and patchy demyelination of peripheral neurons with chromatolysis of anterior horn cells, brain stem nuclei and ganglia of the autonomic nervous system. There may also be diffuse neuronal loss and gliosis of the CNS. The result is damage to autonomic, motor and CNS neurons, giving rise to the characteristic clinical presentation, described below.

Clinical presentation of acute attacks

Acute attacks of porphyria are extremely rare before puberty and unusual after the menopause, having a peak occurrence in the third and fourth decades. Women are more frequently affected than men. In most patients, an

attack will cease once the diagnosis has been made, appropriate treatment given and likely precipitants removed. However, particularly when there is delay in establishing the diagnosis, prolonged and very severe, life-threatening and sometimes fatal acute attacks may occur. Commonly ascribed precipitants include prescribed and illicit drugs, alcohol (particularly binge drinking), systemic infection and dieting either alone or in combination. In women, attacks may be related to the menstrual cycle, usually the luteal phase. However, in some cases no obvious aetiology is found.

An acute attack of porphyria normally starts with continuous abdominal pain, which becomes progressively more severe and is associated with nausea, vomiting and constipation. The patient may also complain of pain in the lower back, buttocks and inner thighs. The abdominal pain is typically diffuse, with no localizing signs or evidence of an acute abdomen on examination. The severity of pain appears out of keeping with the physical signs and often requires the administration of large parenteral doses of opiates. In the absence of an obvious cause, this may trigger concerns about opiate dependence among clinical staff who have no previous experience of the condition. Diminishing pain, particularly where no treatment has been given, does not always herald the end of an acute attack, and patients should be carefully monitored for the development of neuropathy. The autonomic neuropathy that results in the gastrointestinal symptoms also gives rise to tachycardia and hypertension in approximately two-thirds of patients. Patients frequently become dehydrated during an acute attack and may also develop hyponatraemia, which may worsen if inappropriate volumes of hypotonic intravenous fluid are infused. Plasma sodium concentration can fall quickly to extremely low values, precipitating convulsions, which occur in about 5% of patients. Seizures may also occur as a neurological manifestation of the acute attack, and present a particularly difficult management problem as many antiepileptic drugs are implicated as a cause of acute attacks. Rhabdomyolysis, which may lead to renal failure, is a recognized but rare complication of an acute attack.

Severe and untreated attacks can result in a predominantly motor peripheral neuropathy. In most cases, this is a symmetrical distal neuropathy resulting in wrist and foot drop, but occasionally there is progression to a flaccid paresis resembling Guillain-Barré syndrome and requiring prolonged ventilatory support. Prognosis for full recovery is excellent when attacks can be treated and halted. However, patients who require ventilatory support frequently experience a difficult recovery period, characterized by recurrent relapses triggered by the requirement for multiple prescribed drugs and complications such as infection. A small number of patients experience sensory changes such as dysaesthesia or hypoaesthesia in a similar distribution to the motor neuropathy. Other neurological signs reported include cranial nerve involvement including the optic nerve, which can lead to blindness.

The CNS is frequently involved. Acute mental changes are common and include anxiety, insomnia, confusion,

hallucinations and paranoia, which resolve completely following the acute attack. There is no evidence that any form of chronic psychiatric illness is associated with any of the acute porphyrias. Less common CNS features are cerebellar syndrome, pyramidal signs, transient cortical blindness and altered consciousness.

Most patients have only one or a few attacks, with a major attack often being followed by two or three minor ones before full remission. However, about 5% of AIP patients suffer from frequent acute attacks, which in women may be premenstrual, and which may continue for several years (see p. 541). Obvious triggers are usually not found and the repeated admissions to hospital can severely affect the quality of life, particularly if patients have young families. Repeated admissions and high opiate requirement may also lead medical staff who are unfamiliar with the condition to question the diagnosis and the patient's motives.

Chronic complications

Some patients develop a chronic pain syndrome. Pain is usually in the peripheries, often constant and may be triggered by minor stimuli. The cause of this is unknown; management is particularly difficult since it is important to avoid addictive analgesics. Treatment with haem arginate (see p. 540) is usually of no benefit.

Studies from several countries have demonstrated an increased risk of hepatocellular carcinoma (HCC) in the absence of chronic liver disease in patients with an acute hepatic porphyria even when this is clinically latent. The risk is particularly high in Sweden, where a recent study found a standardized increased risk ratio of 64 for AIP gene carriers aged over 54 years and recommended that AIP patients over the age of 50 years should be screened annually for HCC by ultrasound examination. The benefit from screening in other countries has yet to be assessed. It has also been suggested that all HCC patients in whom no obvious cause is found should be screened for acute porphyria.

Renal impairment has been described as a complication of AIP, affecting particularly those patients who have previously suffered acute attacks. Hypertension is also common in these patients, but in many, the decline in renal function precedes the onset of hypertension, which may also be a consequence of the porphyria. Renal biopsies have shown glomerular sclerosis and interstitial fibrosis, but no evidence of inflammation or immunodeposits. A proportion of patients may progress to end-stage renal failure, requiring dialysis and/or renal transplantation. Patients who have had active disease should therefore have their renal function and blood pressure monitored regularly, and antihypertensive therapy should be instituted in an attempt to limit progression of renal impairment.

Diagnosis of acute porphyria

Diagnosis of an acute attack in a newly presenting patient requires the demonstration of increased excretion of urinary PBG in a fresh, random sample of urine that has been protected from light. An assay capable of

quantitative or semiquantitative measurement of PBG should ideally be available in all acute hospitals. In practice, many non-specialist units use rapid qualitative screening tests, particularly out of hours, which have low sensitivity and poor specificity. All positive screening tests should be confirmed by a specific quantitative method with expression of results in relation to creatinine to correct for urine concentration. *If urine PBG and ALA concentrations are normal during the early phase of an acute illness, all acute porphyrias, including ADP, are excluded as a cause of that illness.* However, both PBG and ALA may return to within normal limits within a few days after the onset of symptoms in VP and HCP, and negative findings should be interpreted with caution when there has been a delay in collecting samples. However, in both these disorders, urinary and faecal porphyrins remain elevated for a prolonged period. Conversely, urinary PBG excretion usually remains elevated for many months or even years after an acute attack in AIP, and an increased PBG does not necessarily indicate an acute attack unless a marked increase from baseline can be shown to coincide with symptoms. Management of a clinically diagnosed acute attack should be started immediately, without waiting for confirmation of a positive screening test or determination of the type of acute porphyria.

Establishment of the type of acute porphyria requires analysis of plasma and faecal porphyrins (see Table 28.3; Fig. 28.2). Faecal porphyrin analysis is essential to distinguish HCP, in which coproporphyrin III accounts for most of the increased faecal porphyrin excretion, from AIP, in which faecal porphyrin excretion is normal or only slightly increased without any change in the coproporphyrin isomer ratio. Variegate porphyria can easily be identified by demonstrating a characteristic plasma porphyrin fluorescence emission peak at 625–628 nm. Enzyme analyses and genetic studies are not required for the diagnosis of new cases of clinically overt porphyria. Their use in family studies is detailed later in the chapter.

Laboratory monitoring of patients during and after acute attacks by regular measurement of porphyrin precursors is rarely indicated; treatment should be guided by clinical assessment. The one exception is where a severe attack has progressed to flaccid paresis and clinical assessment is difficult. In these circumstances, weekly monitoring may provide useful information on whether the condition is stable or deteriorating.

Management of an acute attack

Supportive treatment. As soon as the acute attack is confirmed, drugs or other recognized precipitants should be withdrawn. Safe and effective management of symptoms and complications, with support from an expert centre, is essential to minimize the stress; effective pain relief is a major component. This invariably requires opiates, and support from a specialist pain team can be very helpful as very high doses may be required. A phenothiazine may be used for anxiety and restlessness, and to decrease the opiate requirements. Antiemetics should be prescribed for nausea and vomiting. Adequate fluid and energy intake should be ensured, if necessary by intravenous

infusions of 0.9% sodium chloride containing a minimum of 5% dextrose, with regular monitoring of electrolyte status in view of the risk of hyponatraemia. Where drug treatment of precipitants such as infection, coexisting conditions or other features of an acute attack, such as hypertension, tachycardia or convulsions, are required, care should be taken to select medications that are considered safe. Information on drug safety is continually under review and safe drug lists are likely to change regularly (see Welsh Medicines Information Centre: www.wmic.wales.nhs.uk). Where no safe alternative is on the list, an expert centre should be consulted for further advice on patient management. Information about specialist centres offering support in Europe is available from the European Porphyria Network (EPNET): www.porphyrria-europe.org.

Specific treatment. Specific treatment is aimed at suppressing hepatic haem precursor production by administering intravenous haem, which binds to albumin and is transported to the liver, where it downregulates ALAS activity. Treatment is effective if given early, preferably within the first 24–48 h. It will not reverse established neuropathy due to axonal degeneration. Within Europe, haem is available as haem arginate (Normosang® Orphan Europe) and is provided as a concentrated stock solution (25 mg/mL), which should be diluted immediately before use in 100 mL 0.9% saline and administered at a dose of 3 mg/kg body weight in a single intravenous infusion over at least 30 min on each of four consecutive days. As there is no evidence for toxicity, it is practical in most adults to infuse the entire contents of the vial for each dose. However, in smaller patients it may be possible to obtain two doses from each vial, limiting the overall cost of treatment. The main side-effect of therapy is thrombophlebitis at the site of infusion. This can be minimized by careful flushing of the infusion site with normal saline, and re-siting the i.v. cannula after each infusion. Alternatively, the haem can be diluted in 20% human serum albumin, which has proved to be effective in preventing thrombophlebitis, particularly in patients who require regular haem arginate infusions. Lyophilized haem (Panhematin®, Abbott) for intravenous infusion is available in the USA, but is less stable than haem arginate, and has more potential side-effects, including thrombocytopenia and coagulopathy.

With the advent of intravenous haem therapy, carbohydrate loading, which also decreases hepatic ALAS1 activity, is no longer indicated as a specific treatment unless haem is unavailable or there is a delay in obtaining it. High-dose carbohydrate (300–500 g glucose/24 h) should be given intravenously via a central venous catheter.

Preventing acute attacks

All patients should be advised about avoiding factors that increase the risk of an acute attack. These include drugs, alcohol and fasting or dieting, particularly diets in which carbohydrate is avoided completely. Weight reducing diets should be undertaken under the supervision of a dietitian. Patients should also be encouraged to register with an organization such as the MedicAlert Foundation,

which provides jewellery inscribed with appropriate information in case of emergencies. It is helpful to provide patients with written information explaining the disorder and, where applicable, they should be provided with details of existing patient support groups (e.g. the British Porphyria Association: www.porphyrria.org.uk). They should also be offered the opportunity to see a clinician with a special interest in porphyria on at least one occasion. Patients in remission (i.e. those who have had an acute attack) should be followed-up annually, preferably at a specialist centre or in conjunction with one. Local follow-up is also important in order to ensure contact with clinical services should admission be required. Through such shared care arrangements, most intervening problems can be managed successfully through correspondence or over the telephone.

Severely affected patients

A minority of patients with acute porphyria suffer from repeated acute attacks, occasionally as often as every 3–4 weeks. The majority are women with AIP in their 20s and 30s and their management can present a major challenge to clinicians and the acute medical services. A careful history should be taken to ascertain any obvious provoking factor, such as smoking, alcohol, drugs or stage of the menstrual cycle. An individualized management protocol should be drawn up by the local physician with help from a recognized national porphyria service; these operate in many European countries, including the UK, where the National Acute Porphyria Service (NAPS) has been commissioned specifically to support the management of patients with active porphyria. The patient should be encouraged to seek early medical intervention as prompt treatment may abort an acute attack, limiting the length of inpatient stay. This is best achieved by rapid direct access to inpatient facilities and the support of a clinical team who knows the patient and is experienced in dealing with the treatment of acute attacks.

When attacks appear to be associated with the luteal phase of the menstrual cycle, they may be reduced in frequency and/or severity by suppression of ovulation using gonadotrophin releasing hormone agonists. If this succeeds and is to be continued for longer than six months, measures to prevent osteoporosis become essential. Oestrogen supplementation, preferably in the form of patches, is tolerated by some patients, and diminishes any menopausal symptoms, but carries a small risk of provoking further attacks. Progestogens should not be used and most women will therefore require regular endometrial monitoring under the supervision of a gynaecologist. Adequate vitamin D and calcium intake should be ensured, by supplementation if required. Bisphosphonates have also been used in this situation to reduce the risk of osteoporosis.

Regular infusion of haem arginate reduces the frequency and severity of acute attacks in many patients. The minimum effective dose frequency, varying between monthly and weekly single infusions, should be established for each individual patient. There is no typical regimen. Where this treatment is being considered, it is often helpful to contact an expert porphyria centre to discuss the individual details with a clinician with experience

of its use. An indwelling central venous line is invariably required, although the specific type of line used should be dictated by local expertise and patient choice. Dilution of haem arginate in albumin solutions may reduce the need to replace blocked venous catheters.

Where these measures fail to control the attacks with resulting repeated life-threatening crises, problems with central venous access arise, progressive renal dysfunction occurs or quality of life is assessed as very poor, patients should be considered for liver transplantation, which is curative. Careful counselling and assessment is required, preferably in centres with experience with orthotopic liver transplantation (OLT) for this indication. Referral should ideally be before any chronic damage becomes irreversible. Liver transplantation has also confirmed that overproduction of haem precursors by the liver is central to the pathogenesis of acute attacks. This evidence has supported a project to develop gene therapy using a recombinant viral vector targeting the liver as a treatment for AIP.

Recurrent acute abdominal pain should not automatically be ascribed to the porphyria, particularly if the patient reports that the pain is not the same as previously experienced. Other conditions should be considered, as failure to diagnose these can have serious, possibly fatal consequences. In this situation, measuring urine PBG and/or ALA may be helpful in patients with VP or HCP. In AIP patients, a baseline measurement is required to aid interpretation of later results.

Chronic pain, which may be neuropathic, is a common problem in patients who have suffered regular acute attacks. Pain may be in the abdomen, limbs or lower back and may, in some cases, be present almost continuously. Although there may be an initial response to haem arginate, suppression of ovulation or other methods used to manage recurrent acute attacks, this is rarely sustained. Opiates are not a long-term solution in view of the risk of dependence, and should not be prescribed between acute attacks. Analgesia with a non-steroidal anti-inflammatory drug can, in some cases, be successful, and some patients may benefit from medication used to treat neuropathic pain such as gabapentin or pregabalin. Detailed discussion with the patient to explain that these symptoms are not due to an acute attack can be reassuring.

Managing asymptomatic relatives of patients

Once the diagnosis of a specific type of acute porphyria has been established in a new patient, screening should be offered to family members, so that those found to be affected, most of whom will have latent porphyria, can be offered specific advice as to how to limit the risk of suffering an acute attack.

Family studies. Metabolite measurements are highly specific but are almost always normal before puberty and have low sensitivity in adults (Table 28.5). Enzyme measurements are more sensitive but their sensitivity and specificity are limited by the overlap between activities in disease and in normal subjects. They have now largely been replaced by mutation detection by DNA analysis, which is specific and 100% sensitive if the mutation that causes porphyria in the family under investigation is known. It is essential, therefore, that children, and those family members whose specific porphyrin biochemistry is normal, be offered gene testing. Enzyme measurements, for example erythrocyte HMBS assay for detection of latent AIP, and gene tracking using intragenic single nucleotide polymorphisms, may be useful in the few families in which a disease-specific mutation cannot be identified.

Safe prescribing

Prescription drugs are an established precipitant of acute attacks and, even in known porphyria patients, careless prescribing has induced acute attacks or made evolving ones worse. It is therefore essential to provide information concerning the safe use of prescription medication to both those affected and their clinicians. A recent initiative in Europe has resulted in a review of the available evidence on a significant proportion of the current pharmacopoeia and provided an overall assessment of safety for individual drugs (www.drugs-porphyrria.org).

Drugs can be classified into three broad groups, based on clinical experience, experimental evidence and an understanding of their metabolism and excretion. These are: *Not Porphyrinogenic* (safe); *Porphyrinogenic* (unsafe); and *Of Uncertain Safety* (i.e. should be used with caution), which the European system subclassifies into three groups: *Probably Not Porphyrinogenic*, *Possibly Porphyrinogenic* and *Probably Porphyrinogenic*. In the first instance, clinicians should be encouraged to prescribe from a list of drugs that are known to be safe and to avoid those that are definitely unsafe. However, no drug should be seen as completely unusable as there may be clinical situations where a safe alternative is either unsuitable or not available. In these circumstances, an assessment of benefit versus risk should be made, taking into account the severity of the medical condition and porphyria activity. Support from a national centre with expertise in managing porphyria, and access to the most up-to-date drugs information, should also be obtained. If an unsafe drug has to be prescribed, urine PBG should be measured before starting, and at regular intervals during treatment.

TABLE 28.5 Screening asymptomatic relatives for latent acute porphyria

Porphyria	Screening method	Age range (years)	Sensitivity (%)	Specificity (%)
AIP	Urinary PBG >1.5 µmol/mmol	Age ≥15	43	100
HCP	Faecal coproporphyrin III/I molar isomer ratio >1.4	Age ≥7	64	100
VP	Plasma porphyrin fluorescence peak at 624–626 nm	Age ≥15	62	100
AIP, HCP, VP	Detection of known mutation by DNA analysis	All ages	100	100

Specific situations

Pregnancy. Although acute attacks do occur during pregnancy, there does not appear to be a significantly increased risk and most patients tolerate pregnancy, delivery and the puerperium without any adverse consequence. However, it is usually recommended that pregnancy should be delayed until patients have been free of severe acute attacks for a year. Patients should not be allowed to experience prolonged periods of stress and fasting during labour. Effective pain relief, including spinal or epidural anaesthesia, should be administered and, where necessary, intravenous fluids containing glucose should be used to prevent the development of a catabolic state. Where an acute attack does occur during pregnancy, haem preparations have been used and there have been no reports of any adverse effects.

Anaesthesia. General anaesthesia can be safely undertaken in porphyria patients, provided care is taken to select drugs that are known to be safe. General measures to reduce stress and limit periods of fasting should also be instituted, and postoperative complications, such as infection treated aggressively with drugs chosen from the safe list. Regional and dental anaesthesia using local anaesthetics have been safely undertaken in many acute porphyria patients without problem.

Rare forms of acute porphyria

ALA dehydratase deficiency porphyria

5-Aminolaevulinic acid dehydratase deficiency porphyria (ADP) is an autosomal recessive acute porphyria, presenting with neurovisceral symptoms, that results from deficient ALAD activity; the biochemistry is characterized by marked increases in the plasma concentration and urinary excretion of ALA. The prevalence cannot be accurately assessed, but is extremely low with fewer than ten cases reported worldwide to date, and, as yet, none from the UK.

In addition to the very elevated plasma and urinary ALA, other biochemical features include a normal or slightly increased urine PBG, markedly elevated urine coproporphyrin excretion, normal faecal porphyrin and increased erythrocyte protoporphyrin. Confirmation of the diagnosis requires exclusion of lead poisoning, which gives rise to similar biochemical features, and demonstration of decreased ALAD activity in erythroid and non-erythroid cells that is not reversed by addition of excess zinc and a sulphhydryl group reducing agent. Mutational analysis may also be helpful in confirming the diagnosis, and may be required for genetic counselling.

Clinical presentation is variable with reported ages of presentation from birth through to 63 years. The predominant presentation is with acute attacks of abdominal pain and neuropathy that resemble those seen in the autosomal dominant acute porphyrias, and similarly may be triggered by factors such as prescribed drugs that induce ALAS activity. Heterozygote carriers are asymptomatic, but may be at increased risk from the effects of environmental toxins, such as lead, that inhibit ALAD activity. Treatment with hematin and glucose has not been

effective in all cases. Liver transplantation in a patient with ADP did not protect the patient from further acute attacks, nor correct the biochemical abnormalities. The marked enzyme deficiency in other tissues, particularly nervous tissue, probably contributes to the poor clinical outcome in this autosomal recessive condition.

Homozygous acute porphyrias

Homozygosity for null mutations of haem biosynthetic genes is lethal in early embryonic development. However, very rare so-called homozygous variants of all the autosomal dominant acute porphyrias have been reported, in which patients are homozygous or compound heterozygous for mutations that have some residual activity. All present in childhood and have other phenotypic differences from their (heterozygous) autosomal dominant counterparts.

Homozygous AIP is clinically the most severe of these variants and is usually associated with a progressive leukodystrophy. Both homozygous HCP and homozygous VP present with skin lesions in childhood with, in the latter, clinodactyly (curving of the little finger towards the ring finger) and sometimes other abnormalities, including short stature and neurological defects. Acute attacks have been reported in homozygous HCP but not homozygous VP. Missense mutations in exon 6 of the *CPOX* gene cause harderoporphyria, a relatively benign condition characterized by neonatal jaundice, persistent mild haemolytic anaemia, skin lesions and excretion of the tricarboxylic porphyrin, harderoporphyrin, in faeces. All these variants can be distinguished from their (heterozygous) autosomal dominant counterparts by demonstrating very low activity of the relevant enzyme and by mutational analysis. Erythrocyte protoporphyrin concentrations are also increased in all these conditions and this may serve as a useful initial diagnostic indicator. With the exception of harderoporphyria, porphyrin excretion patterns in the homozygous acute porphyrias are indistinguishable from those of the autosomal dominant forms.

THE CUTANEOUS PORPHYRIAS

Bullous porphyrias

The bullous porphyrias all give rise to identical skin lesions, and so cannot be distinguished with certainty on clinical grounds alone. Although patients with CEP or hepatoerythropoietic porphyria (HEP, see p. 545) tend to present in infancy or childhood, and those with PCT, VP or HCP in adulthood, exceptions do occur: occasionally PCT presents in early childhood, while the skin lesions of late onset CEP may first appear in adults and be mistaken for PCT. Biochemical investigation is essential for accurate diagnosis (see Table 28.3). It is important to distinguish between the disorders in view of the differences in prognosis, treatment and susceptibility to acute attacks. In addition, clinically identical skin lesions, termed pseudoporphyria, can occur in the absence of any disturbance of porphyrin metabolism as a consequence of

drug reactions, prolonged use of sun beds and in association with long-term haemodialysis.

Pathophysiology of skin lesions

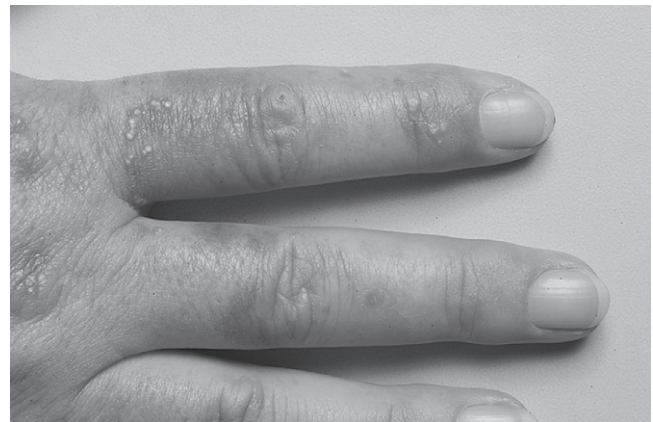
Porphyrin-induced photosensitivity results from the absorption of visible light at the surface of the skin. Porphyrins enter the dermis from the plasma and interact with light of 400–410 nm wavelength, which is capable of penetrating to the level of the dermis and basement membrane. This photoactivation of the porphyrin ring gives rise to an unstable, excited-state molecule, which can release energy as light or react with oxygen by resonance energy transfer to form a singlet oxygen species or superoxide anion. These reactive oxygen species cause cellular damage via reactions with lipids, proteins and DNA, resulting in complement activation, degradation of mast cells and damage to the dermis and basement membrane. This process gives rise to the characteristic features evident on histopathology: perivascular accumulation of amorphous hyaline material involving the small vessels of the dermis. This PAS-positive material includes deposits of immunoglobulin and complement and appears to emanate from the walls and lumens of the blood vessels. Bullae occur where the epidermis splits from chronically damaged, thickened basement membrane and forms a roof for a pocket of clear fluid. The dermal papillae are distorted and flattened by the accumulated amorphous hyaline material and the floors of the bullae have a characteristic festooned appearance.

Skin symptoms and signs

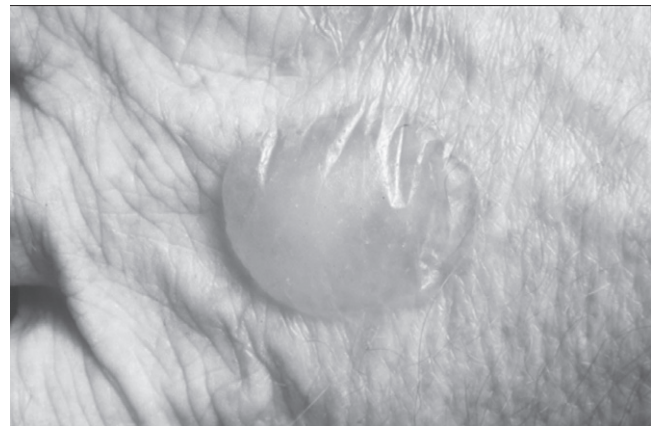
The skin lesions may occur in any sun-exposed area, most frequently the backs of the hands, the forearms, the face and the feet, particularly when open sandals are worn. Fragile skin that tears following minor mechanical stress is the commonest feature. Injury can occur in response to everyday household activities, and patients frequently wear gloves to avoid damaging their skin. Vesicles and bullae filled with clear fluid are commonly present, as are small white spots or milia (Fig. 28.3). Haemorrhagic bullae are uncommon. Bullae rupture easily and form crusted lesions that heal slowly, resulting in chronic scarring of the skin and areas of depigmentation. Secondary infection of open lesions may occur, which worsens the scarring and in severe cases, may result in photomutilation and disfiguration. Other features include hypertrichosis, usually on the forehead and temples; this is particularly noticeable and troubling in women. Less common clinical features include hyper- or hypopigmentary changes, onycholysis, conjunctival damage and scarring alopecia. Because of the chronic nature of the conditions, the relationship to sunlight is not always recognized by the patient.

Biochemical features and diagnostic approach

Each of the bullous porphyrias has a characteristic pattern of porphyrin excretion during active disease that allows the conditions to be distinguished (see



A



B

FIGURE 28.3 ■ The skin lesions associated with bullous photosensitivity. (A) Chronic scarring due to skin fragility, partially healed erosions from ruptured bullae and milia on the back of the hand. (B) A typical clear fluid-filled blister on the sun-exposed back of the hand.

Table 28.3). Ideally, samples of urine, faeces and EDTA-anticoagulated blood should be sent for full porphyrin analysis on all patients. However, the single most useful analysis in patients with active skin lesions is a plasma porphyrin fluorescence emission scan, which, if normal, excludes active cutaneous porphyria (see Fig. 28.2). Many dermatologists now follow this approach and will send the other samples required for diagnosis only if the plasma porphyrin screen is abnormal. However, it is essential that the plasma scan be undertaken in a fluorometer that is sufficiently sensitive, which requires it to be fitted with a red-sensitive photomultiplier. An increased plasma porphyrin fluorescence emission peak with a maximum at 624–627 nm is diagnostic for VP. However, it does not necessarily indicate active porphyria, as positive plasma scans are found in approximately 60% of adults with latent VP and in most of those in remission. An emission maximum at 615–620 nm is consistent with skin lesions caused by porphyria but does not distinguish between PCT, CEP or HCP. However, these conditions can be unequivocally differentiated by urinary and faecal porphyrin analyses.

Individual disorders

Porphyria cutanea tarda. Porphyria cutanea tarda (PCT) is the commonest of the porphyrias, with a prevalence of approximately 1:25 000. It results from partial deficiency of the enzyme uroporphyrinogen decarboxylase (UROD) in the liver and is characterized biochemically by overproduction of uroporphyrin, 7-, 6- and 5-carboxyl porphyrins and isocoporphyrin. In many patients, there is evidence for underlying liver cell damage, and biochemical tests of liver function are often abnormal. Histopathological examination usually reveals only minor abnormalities: mild fatty infiltration with some hepatocyte necrosis and periportal inflammation. Cirrhosis is present in fewer than 20% of patients, but when it does occur, it carries a much greater risk of hepatocellular carcinoma than does comparable cirrhosis without PCT.

The majority of patients in the UK (80%) have sporadic or type I PCT, in which UROD deficiency is restricted to the liver. Most of the remaining 20% have familial or type II PCT, in which UROD deficiency is present in all tissues and is inherited in an autosomal dominant pattern. As with other autosomal dominant porphyrias, the clinical penetrance of familial PCT is low, with symptoms in fewer than 10% of those who inherit the condition. Porphyria cutanea tarda may also be caused by exposure to various halogenated aromatic hydrocarbons as in the outbreak of hexachlorobenzene poisoning in south-eastern Turkey in the late 1950s.

Whether the patient has familial or sporadic PCT, overproduction of sufficient porphyrin to produce symptoms appears to require a decrease in hepatic UROD activity to well below 50% of normal. This is brought about through the reversible inactivation of UROD by a porphomethene inhibitor formed from uroporphyrinogen by an iron-dependent oxidative mechanism, probably catalysed by hepatic cytochrome P450s. Less inactivation of UROD is required for symptoms to occur in familial PCT, where prior enzyme activity is half-normal and, although there is a wide overlap, patients tend to present at a younger age than those with sporadic PCT. Other risk factors for both types of PCT include alcohol misuse, conditions that increase hepatic iron content such as hereditary haemochromatosis, hepatotropic viruses (particularly hepatitis C virus, HCV) and HIV, and prescribed oestrogens: two or more risk factors are often present. Most patients have some hepatic iron overload but amounts are usually below those in overt haemochromatosis. Among patients of northern European descent, about 20% are homozygous for the haemochromatosis (*HFE*) gene C282Y mutation, irrespective of the type of PCT, but heterozygosity for this mutation is a less important risk factor. Prevalences for antibodies to HCV range from 8% to 79%, being highest in southern Europe and the USA and lowest in northern Europe. When PCT develops in HCV, it is usually an early manifestation; in contrast, it occurs as a late complication of HIV infection. Conditions less frequently associated with PCT include chronic renal disease, diabetes mellitus, systemic lupus erythematosus and various haematological malignancies.

Clinical management of patients with PCT should start with the identification of any predisposing factors that can be withdrawn or treated. Newly diagnosed patients should be investigated for underlying liver disease, tested for iron overload by measuring transferrin saturation and *HFE* genotyping, and screened for hepatitis viruses and, if indicated, HIV infection. Where there is evidence of chronic liver disease, referral to a hepatologist is advisable. All patients should be given general advice about avoidance of sunlight, encouraged to protect their hands from trauma by wearing gloves and advised to seek early treatment of any infected lesions. In contrast to patients with acute porphyrias, patients with PCT need not avoid any drugs, apart from antimalarial doses of chloroquine and its derivatives.

Two effective treatments for PCT are available. Venesection of a unit of blood at one- or two-weekly intervals to deplete hepatic iron stores prevents inactivation of hepatic UROD and eventually restores enzyme activity to basal levels. Treatment should be continued until transferrin saturation falls to 16% or less or the patient becomes anaemic (haemoglobin less than 120 g/L). Serum ferritin concentration is less satisfactory for monitoring, as values may be affected by concomitant liver disease. If it is used, treatment should continue until the concentration is 25 µg/L or less. Although resolution of skin lesions occurs within 3–9 months, biochemical remission usually takes longer as large quantities of porphyrins stored in the liver are released and excreted in the urine. Other methods used to reduce iron stores include subcutaneous desferrioxamine and, in patients with renal failure, erythropoietin without iron supplementation. Treatment with low-dose chloroquine (125 mg twice weekly) or hydroxychloroquine (100 mg twice weekly), which complexes with uroporphyrin and mobilizes it for excretion from the liver, is also effective. Larger doses will provoke an acute hepatotoxic reaction with systemic symptoms and should be avoided. Clinical improvement usually becomes evident within 3–4 months and treatment should be continued until total urinary porphyrin excretion falls to normal. Chloroquine treatment may be combined with venesection if response to monotherapy is slow. Choice between these two treatments is often determined by local preference. There is little evidence that chloroquine worsens liver disease, but it is potentially hepatotoxic in PCT and should probably be avoided in patients with severe liver disease. Iron depletion should be used for all *HFE* C282Y homozygotes; it is the only effective method in renal failure and ensures compliance, but is more expensive. Both treatments produce prolonged remission and, particularly in sporadic PCT, may prevent relapse. Nevertheless, long-term management should take into account the possibility of relapse, and patients in remission may benefit from an annual review. Patients with significant liver disease, familial haemochromatosis or other associated conditions should be managed by an appropriate specialist.

Determination of the type of PCT is best carried out by mutational analysis of the *UROD* gene. However, it is not essential for clinical management. Treatment is the same for both types, while the low clinical penetrance, absence of potentially life-threatening symptoms and

availability of effective treatments means that mutational analysis for family studies is difficult to justify, except perhaps for those rare families in which more than one individual has overt PCT.

Hepatoerythropoietic porphyria. This is a rare variant of familial PCT in which patients are homo- or heteroallelic for *UROD* mutations. Uroporphyrinogen decarboxylase activity is markedly decreased in all tissues, although the liver appears to be the main source of excess porphyrins. Skin lesions usually appear in early childhood and are identical to those of PCT. They persist and may occasionally become as severe and disfiguring as in CEP. Other organs are rarely affected; although severe anaemia has been described in HEP, most patients have no significant haematological abnormalities. Differentiation from other early onset porphyrias requires analysis of urinary, faecal and erythrocyte porphyrins, with measurement of erythrocyte *UROD* activity and/or mutational analysis of the *UROD* gene. Photoprotection is the only effective treatment. There is no associated iron overload and patients do not respond to phlebotomy or chloroquine. For most patients, the disease does not appear to affect survival.

Congenital erythropoietic porphyria. Congenital erythropoietic porphyria (CEP), or Gunther disease, is a rare bullous porphyria, with a UK prevalence of 1 in 3 million. It is pan-ethnic, although like other autosomal recessive conditions prevalence is increased in communities where consanguinity is common. Marked deficiency of *UROS* results in the accumulation of large amounts of uroporphyrin I, formed by non-enzymatic cyclization of HMB, and, to a lesser extent, of coproporphyrin I. The main source of excess porphyrin production is the bone marrow, with less than 1% produced from other tissues. The severity of the condition varies from hydrops fetalis, resulting from in utero haemolytic anaemia, to the late onset forms in which the disease becomes manifest in late childhood and young adults. The severity appears to correlate with the level of residual enzyme activity, with the more marked deficiency resulting in very high plasma and red cell porphyrin concentrations, which result in two major pathophysiological consequences: marked skin photosensitivity and haemolytic anaemia with ineffective erythropoiesis and splenomegaly. Most porphyrin is released from cells destroyed in the bone marrow but some comes from red cells sequestered and destroyed by the spleen. The haemolytic anaemia associated with CEP, which may be caused by porphyrin-induced damage to erythroid cell membranes, further stimulates erythropoiesis, increasing porphyrin overproduction.

The majority of patients present soon after birth with skin blisters, occasionally in response to phototherapy treatment for neonatal hyperbilirubinaemia, and red-brown staining of their nappies. The skin lesions are more severe and extensive than in other bullous porphyrias, and progressive damage can lead to photomutilation affecting the ears, nose, eyelids and fingers. Porphyrins also accumulate in bone and teeth, causing discolouration, evident as erythrodontia – reddish-brown teeth that fluoresce red on UV illumination. Bone marrow hyperplasia may lead

to osteolytic lesions. Other skeletal abnormalities include decreased bone density, possibly resulting from low vitamin D due to strict avoidance of sunlight, leading to pathological fractures. The hands are frequently severely affected with contractures, atrophy and bone resorption of the terminal phalanges.

The variable phenotype dictates the treatment approach, and early on in the course of the disease, mutational analysis of the *UROS* gene can provide useful information as there is some correlation between genotype and phenotype. Homozygosity for a mutation present on about 30% of CEP alleles, C73R, carries a particularly poor prognosis. Depending on the combination of mutations, this information can be used to classify the likely clinical severity as mild, moderate or severe and predict likely treatment requirements. Mutational analysis may also be required for prenatal diagnosis.

As with other cutaneous porphyrias, the mainstay of treatment is avoidance of sunlight; special measures such as fitting filters to house and car windows may be required. Aggressive treatment of any infected skin lesions with antibiotics is essential. Regular blood transfusions aimed at maintaining the haematocrit above 33% suppresses erythropoiesis and can successfully decrease porphyrin production and improve photosensitivity. In patients with transfusion-dependent anaemia, it is essential to use desferrioxamine or other iron chelators to control iron accumulation, which has been an important cause of death in the past. If hypersplenism occurs, splenectomy may reduce haemolysis and transfusion requirements. Other treatment modalities reported on small numbers or individual patients, and shown to be of limited benefit, include β -carotene, oral activated charcoal, plasmatorbent therapy, hydroxyurea and haematin infusion. Bone disease can be managed with bisphosphonate therapy and vitamin D supplementation where deficiency is evident. Prophylactic vitamin D should also be considered.

Allogeneic bone marrow transplantation (BMT) has been undertaken successfully in more than ten patients with CEP, resulting in a marked decrease in porphyrin production and photosensitivity in all cases. The main limitation is the availability of suitable donors. Use of cells from a heterozygous donor does not appear to affect the outcome. The major risks are the morbidity and mortality associated with the procedure itself, and bone marrow transplantation should therefore be reserved for the most severely affected patients. Clinical care for patients with CEP should be provided by a multidisciplinary team, including a dermatologist, haematologist, ophthalmologist and a porphyria expert, in order to assist patients and their families make informed decisions about this form of treatment.

Gene therapy for the erythropoietic porphyrias, particularly CEP, is being actively pursued. The success of BMT indicates that introduction of active enzyme into a sufficient number of deficient cells by ex vivo transformation should be curative. Gene delivery methodology has been successfully developed for other bone marrow disorders, for example X-linked severe combined immunodeficiency.

Erythropoietic protoporphyria and X-linked dominant protoporphyria

Two clinically indistinguishable porphyrias present with acute photosensitivity without skin fragility. In both, the photosensitivity results from accumulation of protoporphyrin in the skin. The concentration of protoporphyrin is also increased in erythrocytes, plasma, liver and other tissues. Most patients with this presentation have erythropoietic protoporphyria (EPP) but about 5% have XLDPP.

The excess protoporphyrin is produced mainly or, in XLDPP, exclusively by erythroid cells in bone marrow. In EPP, protoporphyrin accumulation results from a decrease in ferrochelatase (FECH) activity to less than 35% of normal. Almost all patients are compound heterozygotes for one of a large number of *FECH* mutations that abolish or severely decrease FECH activity and a hypomorphic allele (*FECH* IVS3–48C) that has about three-quarters of the activity of the normal allele and is present in around 10% of white Europeans, 30–45% of Asians and <3% of Africans. About 3% of patients lack the hypomorphic allele but have *FECH* mutations on both alleles, at least one of which retains some FECH activity. Within families, EPP segregates as an autosomal recessive disorder that may show pseudo-dominant inheritance in populations where the frequency of the hypomorphic allele is high. A late onset form of EPP associated with myelodysplasia or myeloproliferative disease and caused by acquired somatic *FECH* mutation has also been described.

X-linked dominant protoporphyria is caused by mutations in *ALAS2* that disrupt the C-terminal region of erythroid *ALAS2*, thereby increasing its activity and leading to increased formation of protoporphyrin. Unlike in EPP, FECH activity is normal, which enables conversion of some of the protoporphyrin that accumulates to zinc-protoporphyrin. Families normally show inheritance through several generations. This pattern is unusual in EPP and is an important indicator of the possible presence of XLDPP. As might be expected for a gain in function mutation on an X chromosome, both sexes are similarly affected, though skewed inactivation of the mutant chromosome may lead to mild or no disease in a small number of women.

The diagnosis of both conditions can be established by demonstrating an increase in erythrocyte protoporphyrin using a method that distinguishes between free protoporphyrin and its zinc chelate, which is the predominant form in normal and iron-deficient erythrocytes. In EPP, almost all the protoporphyrin is in the free form; more than 15% of zinc-protoporphyrin in a patient with a markedly increased total protoporphyrin indicates possible XLDPP, in which zinc-protoporphyrin forms from 20% to 70% of the total (see Fig. 28.2). Mutational analysis of *ALAS2* is required to confirm a diagnosis of XLDPP. Protoporphyrin is increased in plasma in EPP and XLDPP and has a fluorescence emission peak at 626–635 nm. Faecal protoporphyrin excretion is increased in about 40% of EPP patients, and some of those with XLDPP, but is not specific for either condition and is of limited value diagnostically. Protoporphyrin is hydrophobic and is not excreted in urine, which therefore cannot be used to diagnose or exclude EPP or XLDPP.

Skin symptoms and signs

Both conditions are characterized by life-long acute photosensitivity with pain and swelling of sun-exposed skin, usually becoming evident within the first two years of life. Although symptoms typically appear in early childhood, diagnosis may be delayed for many years unless there is a family history. Patients give a characteristic history of burning pain within minutes of exposure to sunlight, which may only be relieved by cold water or wet towels. There may also be priming phenomena, whereby exposure to sunlight sensitizes the skin, lowering tolerance to sunlight on subsequent occasions. The oedema evident immediately after exposure has usually resolved by the time a doctor sees the child, hence the difficulty in diagnosis. With prolonged exposure, an erythematous reaction or even blistering can occur. Over time, patients develop chronic skin lesions with a waxy, thickened appearance. Chronic exposure of the face leads to linear scarring of the forehead and cheeks, with furrowing around the mouth (Fig. 28.4). Occasionally, marked thickening of the skin develops over the joints of the hands and fingers.



A



B

FIGURE 28.4 ■ The skin lesions associated with chronic sunlight exposure in erythropoietic protoporphyria. (A) Linear and pitted scarring of the forehead. (B) The back of the hands and fingers showing marked thickening of the skin with a waxy, lichenoid appearance.

Severe acute photosensitivity has a serious adverse effect on quality of life, sometimes worsened by delayed or incorrect diagnosis. Childhood can be a particularly difficult period and adult patients frequently state that they do not want to have children if they have to go through a similar experience.

Treatment

Avoidance of sunlight is the mainstay of management of acute protoporphyrin-induced photosensitivity. Outdoor clothing should include full-length garments to cover the arms and legs, a wide brimmed hat (or peaked hat with a cloth covering the neck), closed shoes and gloves if possible. Sunscreen creams must be formulated to block wavelengths around 400–420 nm. The most effective are those based on titanium dioxide, which acts as a reflectant, and a specially formulated product available in three tinted versions (coral, beige and coffee) is available in the UK (Tayside Pharmaceuticals, Ninewells, Dundee). Where prevention fails, cold compresses and application of local corticosteroid creams to the affected area can provide symptomatic relief. Antihistamines may also provide some relief by reducing the immediate release of histamine by mast cells.

Several specific treatments have been proposed, although none has proved to be completely reliable. The most promising is narrow-band UV therapy, which induces sunlight tolerance equivalent to a sun protection factor (SPF) of 8. Effective treatment is limited by availability of equipment and the requirement for regular therapy. More traditionally, patients should be offered a trial of β -carotene therapy, which is believed to work by quenching oxygen radicals and/or interfering with the absorbance of light by protoporphyrin. Efficacy is variable, with a significant proportion of patients reporting no benefit. However, it is important to ensure an adequate dose of between 100 and 300 mg/day, sufficient to achieve plasma concentrations of 11–15 $\mu\text{mol/L}$ (6–8 mg/L). Hormonal tanning agents based on α -melanocyte stimulating hormone (α -MSH) analogues are currently being investigated as a treatment and have shown promise in early trials.

Bone marrow transplantation (BMT) has also been used successfully in several EPP patients, either to treat a coexistent haematological malignancy or to protect a transplanted liver from recurrence of protoporphyrin hepatopathy (see below). Although in most EPP patients, the risk of BMT outweighs the benefits, if it becomes possible to identify patients with a high risk of protoporphyrin liver disease, BMT could be used to cure the condition and eliminate the need for liver transplantation.

Chronic complications and their management

Mild microcytic, normochromic anaemia, which is associated with reduced iron indices, is common in both conditions, affecting up to a third of EPP patients. Bone marrow studies show scattered sideroblasts and perimitochondrial iron accumulation, suggesting defective utilization of iron for haem synthesis. Iron therapy is, therefore, unlikely to resolve the anaemia and, in some cases, has been reported to worsen photosensitivity.

The most serious complication of EPP and XLDP is progressive protoporphyrin liver disease, which affects 1–2% of patients and may lead to severe acute liver failure. Excess protoporphyrin is excreted via the biliary tract and deposits of protoporphyrin are present in the hepatocytes of many patients. In a minority, this leads to liver damage that, although initially mild as indicated by small increases in plasma transaminase activities, will in some cases progress to cirrhosis. Further deterioration leads to cholestatic jaundice and abdominal pain that may be similar to the neurovisceral pain experienced in the acute porphyrias. Hypersplenism and the accompanying haemolysis stimulate erythropoiesis and, in combination with the decreased biliary protoporphyrin excretion, result in increased red blood cell protoporphyrin concentrations and worsening photosensitivity. This fulminant or decompensated liver failure is usually fatal unless the patient receives a liver transplant. Perioperative risks include severe phototoxic burns as a result of high levels of circulating protoporphyrin. This can be reduced by pre-operative treatments that reduce protoporphyrin concentrations; it is essential that the theatre lights be filtered. In addition, patients are at risk of neuronal damage leading to abdominal pain and profound motor neuropathy requiring ventilation. Finally, protoporphyrin liver disease can recur in the transplanted liver and, in these circumstances, bone marrow transplantation to correct the metabolic defect and prevent further hepatic damage is the only viable option.

There is no reliable way of predicting which patients are at risk of severe liver disease, although there appears to be an increased risk in patients with XLDP and those with true autosomal recessive EPP (mutations on both *FECH* alleles). All patients should therefore be seen regularly throughout their lives by a porphyria specialist to monitor both erythrocyte protoporphyrin concentrations and liver function. Patients in whom liver function is found to be abnormal, or whose protoporphyrin concentrations are markedly increased, should be referred to a hepatologist for full assessment and liver biopsy. At this stage, treatment to halt the progressive liver damage caused by the increasing protoporphyrin concentrations should be instituted. This should include avoiding alcohol, measures to enhance protoporphyrin excretion such as oral cholestyramine, activated charcoal or bile acids (ursodeoxycholic or chenodeoxycholic acid). In patients with incipient liver failure, particularly those awaiting liver transplantation, this should be combined with measures to reduce protoporphyrin production using hypertransfusion or haematin infusion to suppress erythropoiesis. Finally, exchange transfusion and plasmapheresis can be used to reduce circulating protoporphyrin concentrations immediately prior to transplantation.

SECONDARY DISORDERS OF PORPHYRIN METABOLISM

Abnormalities of porphyrin production and excretion may occur as a secondary manifestation of a large number of disorders other than the porphyrias (Table 28.6). These secondary disturbances of porphyrin metabolism

TABLE 28.6 Secondary disorders of porphyrin metabolism

Porphyrin abnormalities	Condition	Differentiation from porphyria
Increased coproporphyrin I and III in urine	Hepatobiliary dysfunction Chronic liver disease Alcohol, drugs Severe illness, infections	Normal urinary PBG/ALA Normal plasma porphyrin Normal faecal porphyrins
Increased dicarboxylic porphyrins in faeces	Increased haem in gut: diet or bleeding	Normal faecal coproporphyrin Normal plasma and erythrocyte porphyrin Normal urine porphyrins
Increased erythrocyte zinc protoporphyrin	Iron deficiency Haemolytic and other anaemias Lead poisoning	Normal or slightly increased erythrocyte free protoporphyrin Normal plasma porphyrin
Increased urinary ALA and coproporphyrin III	Lead poisoning	PBG normal or increased much less than ALA Increased blood lead
Increased plasma porphyrin	Chronic renal failure (particularly long-term haemodialysis) Cholestatic jaundice	Normal faecal porphyrins

are usually mild, but together are far more common than the inherited porphyrias and may cause clinical and diagnostic confusion unless recognized for what they are.

Urinary coproporphyrin excretion may be increased in any condition that impairs hepatobiliary excretion, by ingestion of alcohol and drugs that induce cytochrome P450 isoforms, in poisoning by lead, other heavy metals and halogenated hydrocarbons, and in many severe illnesses, including infections. In normal individuals, coproporphyrin I predominates in bile, while the III isomer is excreted mainly in urine. As biliary excretion declines, the proportion of coproporphyrin I increases in urine. In alcoholism uncomplicated by liver disease, and in lead poisoning, urinary excretion of coproporphyrin III is increased. In Dubin–Johnson syndrome, coproporphyrin I comprises 80% or more of the total urinary coproporphyrin excretion, which is usually normal or near normal: this pattern has been regarded as diagnostic and distinct from the patterns found in Rotor syndrome and primary biliary cholestasis, which resemble those of other hepatobiliary disorders.

When total urinary porphyrin alone is measured, secondary coproporphyrinuria may suggest the presence of a porphyria but can readily be distinguished by demonstrating normal urinary PBG and ALA and faecal porphyrin excretion, with measurement of blood lead if increased coproporphyrin III or ALA suggests lead poisoning.

Haem is metabolized by the gut flora to protoporphyrin and related dicarboxylic porphyrins. Thus, even small increases in the haem content of the gut, which may not be large enough or occur sufficiently distally in the gut to produce a positive test for occult blood, may increase the total porphyrin content of faeces. Fractionation by HPLC shows that the increase is caused entirely by dicarboxylic porphyrins. Provided a dietary source of extra haem can be excluded, this finding may indicate alimentary tract bleeding of pathological significance.

Erythrocyte porphyrin concentrations are increased in iron deficiency, some other anaemias and in lead poisoning. In these conditions, the erythrocyte porphyrin is almost entirely zinc protoporphyrin and plasma porphyrin

concentrations are normal. In some patients with iron deficiency, faecal porphyrin content may also be increased by gastrointestinal bleeding.

Plasma porphyrin concentrations are increased in cholestatic jaundice and in chronic kidney disease. Patients on long-term haemodialysis may develop a bullous dermatosis on sun-exposed skin that resembles PCT. Since PCT is also associated with end-stage kidney disease, and can be successfully treated by decreasing hepatic iron stores, it is important to distinguish between the two conditions. Plasma uroporphyrins and heptacarboxylic porphyrins are increased in both, though usually to a greater extent in PCT. Faecal isocoproporphyrin is increased only in PCT and its measurement is more reliable than plasma porphyrin analysis for identifying PCT in such patients.

CONCLUSION

The porphyrias are metabolic diseases characterized biochemically by the excessive production of porphyrins and their precursors as a result of defects in the enzymes leading to haem synthesis. Decreased haem synthesis leads to increased activity of the rate-limiting enzyme, 5-aminolaevulinic acid synthetase and accumulation of the substances involved in the pathway before the block. Porphyrias can be classified as acute (characterized by attacks of a predominantly neurovisceral nature) and chronic (characterized by photosensitivity). Some exhibit both characteristics. They can be distinguished by the pattern of porphyrin excretion, which is unique to each condition.

Further reading

Anderson KE, Sassa S, Bishop DF et al. Disorders of heme biosynthesis: X-linked sideroblastic anemia and the porphyrias. In: Scriver CR, Beaudet AL, Sly WS et al. editors. *The metabolic and molecular basis of inherited disease*. 8th ed. New York: McGraw-Hill; 2001. p. 2961–3062.

A complete and readable chapter on the porphyrias in the definitive textbook on inherited metabolic disease.

Badminton MN, Whatley SD, Deacon AC et al. The porphyrias and other disorders of porphyrin metabolism. In: Burtis CA, Ashwood ER,

- Bruns DE, editors. Tietz textbook of clinical chemistry and molecular diagnostics. 5th ed. St Louis: Elsevier Saunders; 2012. p. 1031–55.
This chapter, in a well known clinical chemistry textbook, has an emphasis on laboratory methodology.
- Deacon AC, Elder GH. ACP Best Practice No. 165: front line tests for the investigation of suspected Porphyria. *J Clin Pathol* 2001;54:500–7.
Detailed information on laboratory methodology used for investigation of porphyria in non-specialist laboratories.
- Desnick RJ, Astrin KH. Congenital erythropoietic porphyria: advances in pathogenesis and treatment. *Br J Haematol* 2002;117:779–95.
Review of CEP with a detailed section on molecular genetics, genotype phenotype correlations and treatment.
- Deybach JC, Puy H. Hepatocellular carcinoma without cirrhosis: think acute hepatic porphyrias and vice versa. *J Intern Med* 2011;269:521–4.
Overview of knowledge and understanding of the current situation regarding this complication of the acute porphyria, particularly AIP.
- Gouya L, Martin-Schmitt C, Robreau A-M et al. Contribution of a common single-nucleotide polymorphism to the genetic predisposition for erythropoietic protoporphyria. *Am J Hum Genet* 2006;78:2–14.
Description of the molecular mechanism underlying penetrance in EPP, the contribution and phylogenetic origin of this polymorphism and its contribution in different population groups.
- Hift RJ, Meissner PN. An analysis of 112 acute porphyric attacks in Cape Town, South Africa: evidence that acute intermittent porphyria and variegate porphyria differ in susceptibility and severity. *Medicine (Baltimore)* 2005;84:48–60.
A clinical description and detailed evaluation of acute attacks from first hand experience.
- Hift RJ, Peters TJ, Meissner PN. A review of the clinical presentation, natural history and inheritance of variegate porphyria: its implausibility as the source of the 'Royal Malady'. *J Clin Pathol* 2012;65:200–5.
Authoritative review of the evidence for this historical diagnosis based on extensive experience of variegate porphyria in South Africa.
- Meyer UA, Schuurmans MM, Lindberg RL. Acute porphyrias: pathogenesis of neurological manifestations. *Semin Liver Dis* 1998;18:43–52.
Critical review of the experimental data for the various hypotheses for neurological dysfunction in the acute porphyrias and the information from experiments in the mouse model for AIP.
- Puy H, Gouya L, Deybach J-C. Porphyrias. *Lancet* 2010;375:924–37.
Recent comprehensive review from one of the leading porphyria centres in Europe.
- Sarkany RP. The management of porphyria cutanea tarda. *Clin Exp Dermatol* 2001;26:225–32.
Useful review of the treatment options for patients with PCT.
- Stein P, Badminton M, Barth J et al. Best practice guidelines in the management of patients with inherited porphyria and their complications. *Ann Clin Biochem* 2013;50:217–23.
Consensus view on managing acute attacks of porphyria written by the clinicians providing the national clinical service for severe acute porphyria in the UK.
- Whitley SD, Badminton M. The role of genetic testing in the management of patients with inherited porphyria and their families. *Ann Clin Biochem* 2013;50:204–16.
Review dealing with the role of genetics in managing porphyria.
- Whitley SD, Ducamp S, Gouya L et al. C-Terminal deletions in the ALAS2 gene lead to gain of function and cause a previously undefined type of human porphyria, X-linked dominant protoporphyria, without anemia or iron overload. *Am J Hum Genet* 2008;83:408–14.
First description and characterization of this new porphyria.

Internet resources

- British Porphyria Association. www.porphyrria.org.uk.
British patient support group website with links to other patient associations.
- Cardiff Porphyria Service. www.cardiff-porphyrria.org.uk.
Website of the authors' department with information about their services and a link to annually updated safe drug list.
- The European Porphyria Network (EPNET). www.porphyrria-europe.org.
Website with details of specialist porphyria centres in Europe, information for patients in 12 different languages, information on diagnosis and treatment and regularly updated information on safe prescribing in the acute porphyrias.
- University of Cape Town Porphyria Service. www.porphyrria.uct.ac.za.
Website of an expert group dealing with all porphyrias, with information for patients and professionals. Particularly useful section on patient management.

The haemoglobinopathies

David C. Rees • Roopen Arya

CHAPTER OUTLINE

INTRODUCTION 550

The structure and function of haemoglobin 550

The genetic control of haemoglobin synthesis 551

THE THALASSAEMIAS 552

α Thalassaemia 552

β Thalassaemia 553

STRUCTURAL HAEMOGLOBIN VARIANTS 554

Sickle cell anaemia 555

Other structural haemoglobin variants 556

LABORATORY DIAGNOSIS OF HAEMOGLOBINOPATHIES 557

CONCLUSION 559

INTRODUCTION

Since the discovery of its central role in oxygen transport in the 19th century, haemoglobin has proved to be a source of fascination to scientists and physicians alike, and is arguably the most widely studied protein in man. Its principal disorders, the thalassaemias and sickle cell disease, quantitative and qualitative defects of synthesis, respectively, are among the most common serious inherited diseases in man and have been pivotal in efforts to understand the molecular basis of human diseases.

It is estimated that about 7% of the world's population carry the mutation for at least one type of haemoglobinopathy, and that 400 000 babies are born each year with a severe form. Approximately 230 000 babies are born each year with sickle cell disease in sub-Saharan Africa, and 120 000 are born with severe thalassaemia in southern and South-East Asia. There is good epidemiological and experimental evidence that the extraordinarily high frequency of haemoglobinopathies and their geographical distribution are due to the relative protection that they provide against malaria and its serious complications. This principally involves improved survival and reproductive success for heterozygotes compared with the normal population, which outweighs the reduced reproductive success of homozygotes. Thalassaemia reaches polymorphic frequencies (>1%) in nearly all populations, except those of northern Europe, northern Asia and the indigenous peoples of Australia, the Americas and the Arctic. The common abnormal haemoglobins have more limited distributions (Fig. 29.1).

Since there is considerable overlap in the world distribution of thalassaemias as well as the common structural haemoglobin variants, co-inheritance of more than one haemoglobin abnormality is common. This generates an extremely diverse range of clinical phenotypes.

The structure and function of haemoglobin

Haemoglobin is a tetramer comprising two α and two β globin chains ($\alpha_2\beta_2$). One molecule of haem, which contains iron and can bind to oxygen, is attached to each globin chain, lying in a hydrophobic cleft. The amino acid sequence of this haem pocket shows marked homology between different animal species, suggesting a specific and important function. Through its iron atom, each haem group is capable of binding one molecule of oxygen. The reversible binding of oxygen to haemoglobin is allosteric, giving rise to the characteristic shape of the oxygen dissociation curve (Fig. 29.2). On binding oxygen, conformational changes occur in both individual globin chains and the tetrameric structure of haemoglobin. The predominant change is closure of the gap between the two β -chains. The site at which this occurs is of critical functional importance. 2,3-Diphosphoglycerate (2,3-DPG), the principal regulator of oxygen affinity in red cells, binds to this part of the haemoglobin molecule. This serves to separate the two β -chains, favouring the deoxygenated conformation and thereby reducing the oxygen affinity of haemoglobin. The observation that oxygen affinity is often reduced in anaemic patients is explained by the adaptive increase in intracellular 2,3-DPG concentration. This enhances oxygen delivery to the tissues and is one of the important reasons why patients frequently tolerate chronic anaemia with very few symptoms. Conversely, the higher oxygen affinity of fetal blood necessary to maintain adequate maternal–fetal oxygen transport is, in large measure, due to the lower affinity of 2,3-DPG for fetal haemoglobin (HbF).

The other main physiological influence on haemoglobin oxygen affinity is the Bohr effect. First recognized at the turn of the century through the effect of carbon dioxide on lowering the oxygen affinity of whole blood, it is now known to reflect the sensitivity of oxygen binding by haemoglobin to changes in blood pH. An increase in

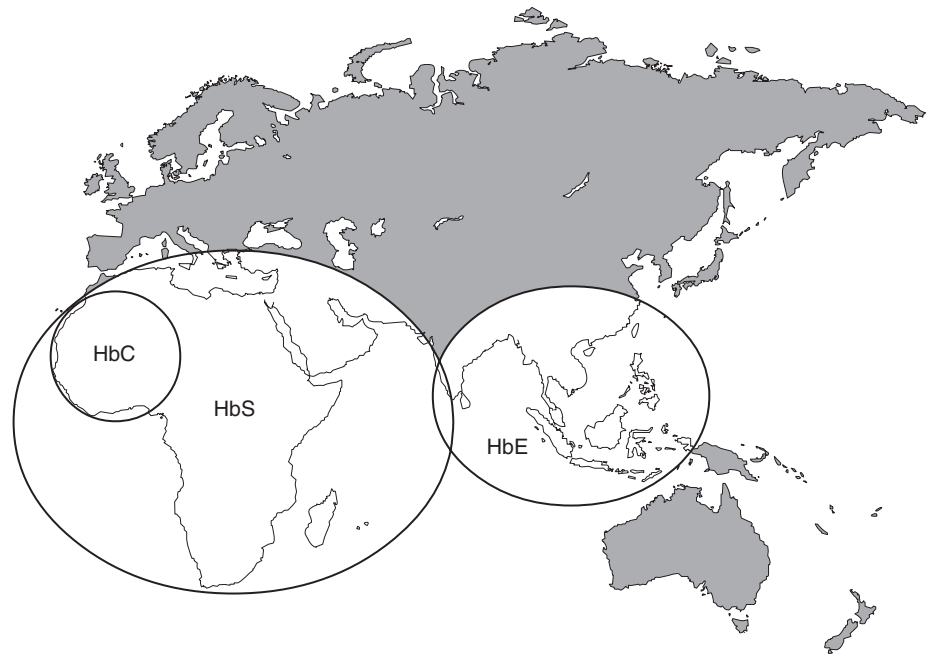


FIGURE 29.1 ■ Primary geographic distribution of haemoglobins (Hb) S, C and E.

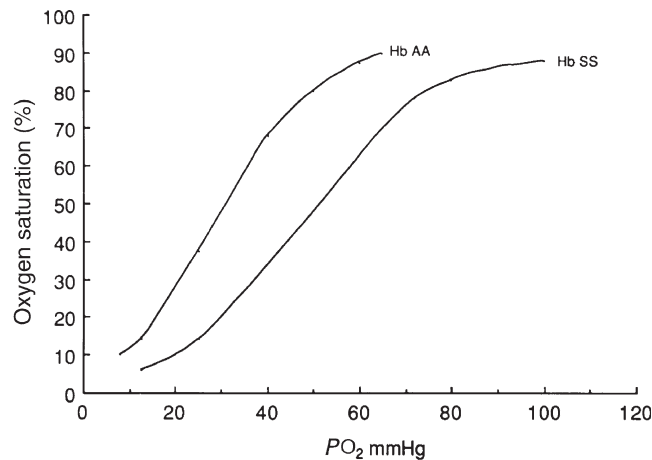


FIGURE 29.2 ■ The oxygen dissociation curves of normal (AA) and sickle (SS) haemoglobin (Hb).

hydrogen ion concentration stabilizes the deoxygenated conformation of the haemoglobin molecule, and so physiological acidosis leads to a reduction in the oxygen affinity. If acidosis is sustained, this effect may be counterbalanced by the reduction in intracellular 2,3-DPG concentration due to modulation of red cell glycolysis. Several inherited variants of haemoglobin have been described that stabilize the molecule in either the oxy or deoxy conformation. In some cases, the interaction with 2,3-DPG or the Bohr effect is modified.

A further factor affecting haemoglobin function is temperature. An increase in temperature reduces oxygen affinity, whereas a fall increases it. As the human is normally isothermic, this is usually of little consequence in terms of physiological adaptation, though during exercise, when oxygen consumption is increased, a raised temperature in the muscles together with an increase in hydrogen ion concentration favours release of oxygen by

haemoglobin. Extreme hypothermia has many adverse consequences, including increased haemoglobin oxygen affinity causing reduced oxygen delivery to tissues.

The genetic control of haemoglobin synthesis

Human globins are encoded by the α and β gene clusters located on chromosomes 16 and 11, respectively (see Fig. 29.3). The α -like genes include an embryonic gene (ζ) and two adult genes (α_1 and α_2). The β -like genes comprise an embryonic gene (ϵ), duplicated fetal (γ) and adult (β and δ) genes. Expression of these genes is developmentally regulated during fetal life. The array of α - and β -like genes on their respective chromosomes reflects the order in which they are expressed. The control of the switch from embryonic, to fetal and then to adult globin synthesis, seems to be controlled by the interaction between several transcription factors, the locus control regions upstream of the gene cluster and the promoter regions of the individual genes.

Normally, the synthesis of α - and β -like chains is carefully balanced to avoid the accumulation of free globin chains, although it is not clear how this balance is maintained. In the thalassaemias, where this balance is severely disturbed, the excess of one type of globin chain is central to the pathophysiology.

The switch in globin synthesis during development has important implications for clinical expression of the haemoglobinopathies. Complete failure of α -chain synthesis, which, since the α genes are duplicated on each chromosome, occurs only when there is loss of function involving all four α genes, becomes evident early in fetal life. By contrast, even if there is little or no normal β -chain synthesis, the effects will not be manifest until the switch from fetal γ to adult β gene expression is completed after birth. This switch is gradual, and abnormalities of β globin synthesis rarely result in clinical problems

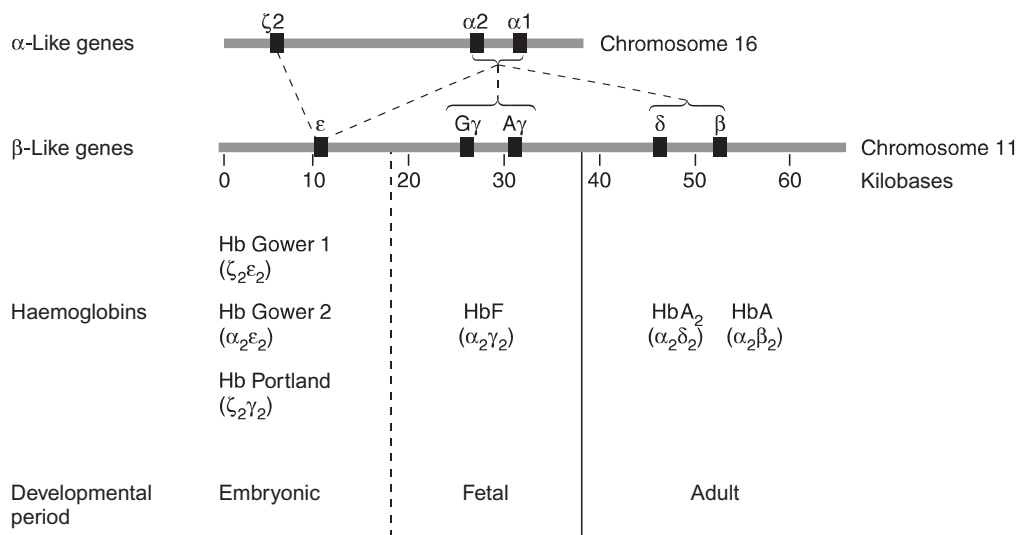


FIGURE 29.3 ■ Organization of the human globin gene cluster and developmental changes in haemoglobin.

before three months of age. Reversing this switch, and reactivating fetal haemoglobin is also potentially curative for diseases caused by mutations of the β globin genes.

THE THALASSAEMIAS

The thalassaemias are one of the commonest human autosomal recessive disorders, with approximately 120 000 severely affected individuals born annually worldwide. The most common and clinically significant forms are α and β thalassaemia and the β thalassaemia-like structural variant, haemoglobin E. These may be further subdivided according to whether there is no output of globin (α^0 or β^0 thalassaemia) or some preservation of globin chain synthesis (α^+ and β^+ thalassaemia). In the case of α thalassaemia, this is complicated further by the duplication of expressed α genes. α Thalassaemia is clinically significant only when three or four α globin genes are lost. The clinical consequences of thalassaemia in the homozygous (or compound heterozygous) state may be understood in terms of the imbalance in globin chain synthesis that results from absent or reduced synthesis of either α or β globin. Unpaired α or β globin chains are toxic, and damage the developing red cell, causing it to die whilst still in the marrow, which is characteristic of thalassaemia and called ineffective erythropoiesis. Heterozygotes are

unaffected clinically, and manifest the condition as slight anaemia with reduced red cell size and haemoglobin content. Increasing globin chain imbalance, as occurs in homozygous and compound heterozygous states, results in more marked anaemia and bone marrow expansion, with corresponding development of symptoms.

A characteristic of thalassaemia syndromes is their marked phenotypic heterogeneity. For over 20 years now, the molecular basis of thalassaemia has been studied and understood in great detail. The number of mutations identified as causing thalassaemia is large and continues to grow. However, in most populations, a small range of 5–10 different thalassaemia mutations accounts for about 90% of the mutant alleles found in that particular area.

α Thalassaemia

Though particularly common in South-East Asia, where carrier rates may reach 50%, α thalassaemia is widely distributed in all major populations apart from northern Europeans (see Fig. 29.1). The pathophysiological effects of α thalassaemia reflect the degree of impairment in α globin production (see Table 29.1). The majority of cases are due to deletions involving one or both α globin genes.

Haemoglobin Bart's hydrops fetalis is the most severe form of α thalassaemia, where all four α genes are affected, abolishing or severely diminished α chain synthesis.

TABLE 29.1 The α thalassaemia syndromes

Phenotype	Number of functional α genes	Genotype ^a	%Hb Bart's at birth	HbH inclusions
Normal	4	$\alpha\alpha/\alpha\alpha$	0	None
α Thalassaemia trait (minor)	3	$-\alpha/\alpha\alpha$	0–2	None
	2	$-\alpha/-\alpha$ or $---/\alpha\alpha$	2–8	Occasional
HbH disease	1	$---/-\alpha$	10–40	Numerous
Hb Bart's hydrops fetalis	0	$---/---$	80	Present in some cases

^a $-\alpha$, denotes loss of function of one α gene on a chromosome, i.e. α^+ thalassaemia.

$---$, denotes loss of function of both α genes on a chromosome, i.e. α^0 thalassaemia.

The disease becomes manifest in fetal life. With the failure of normal production of fetal haemoglobin ($\alpha_2\gamma_2$), surplus fetal γ -chains combine to form tetramers (γ_4) recognized electrophoretically as haemoglobin Bart's. The fetus is only able to survive at all in utero because of the presence of increased amounts of the embryonic haemoglobin Hb Portland ($\zeta_2\gamma_2$). Haemoglobin Bart's possesses markedly increased oxygen affinity and is ineffective as an oxygen transporter. In an attempt to compensate for impaired tissue oxygen delivery, there is erythroblastosis with extramedullary haemopoiesis, leading to hepatic and splenic enlargement. Severe functional anaemia leads to tissue hypoxia, increased capillary permeability, and cardiac failure, which lead ultimately to fetal hydrops and death in utero or stillbirth in late pregnancy.

Haemoglobin H disease occurs when there is loss of function of three out of four α genes. There is sufficient residual α -chain synthesis to allow production of some normal fetal and adult haemoglobin, and fetal development is generally normal. A variable amount (10–40%) of Hb Bart's is detectable at birth. After birth, the excess β -chains form tetramers detectable electrophoretically as the fast variant haemoglobin H (β_4). Haemoglobin H (HbH) precipitates within red cells with the formation of inclusion bodies leading to shortened red cell survival. Staining of these inclusions with the redox dye brilliant cresyl blue serves as a useful diagnostic test. The clinical effects of HbH disease vary considerably. Most patients have a mild to moderate haemolytic anaemia, which may be exacerbated during pregnancy or parvovirus B19 infection, accompanied by splenomegaly. Growth and development are usually normal. As in other congenital haemolytic anaemias, there is an increased tendency to form pigment gallstones.

It follows that Hb Bart's hydrops fetalis and HbH disease occur only when at least one parent carries the α^0 genotype ($-/\alpha^0$). This provides an explanation for the observation that these disorders occur primarily in southern China, South-East Asia and the eastern Mediterranean where α^0 thalassaemia is prevalent, and are not seen in other parts of the world, for example Africa and India, where α^0 thalassaemia is very rare.

Individuals in whom one or two α genes are dysfunctional are clinically unaffected, although they show thalassaemic red cell indices and traces of Hb Bart's at birth (see Table 29.1). This form of α thalassaemia trait is very common in most populations, with a 3.7 kb deletion accounting for most cases. Non-deletional forms of α thalassaemia are increasingly recognized as DNA analysis becomes more widespread and sophisticated.

β Thalassaemia

β Thalassaemia is prevalent throughout tropical and North Africa, the Mediterranean, Middle East and large parts of south and South-East Asia, including India (see Fig. 29.1). In some parts of the world, for example Cyprus, the heterozygote (carrier) frequency may reach 15%. Over 300 different mutations of the β globin gene or its promoter are known to cause β thalassaemia, the majority being single nucleotide substitutions (point mutations) or small deletions. Geographically, these

segregate non-randomly, so that four or five specific mutations account for the majority of β thalassaemia genes in individual ethnic or geographic groups. Each mutation typically occurs on a single β globin gene haplotype, suggesting that in most instances individual β thalassaemia mutations have arisen historically on a single occasion.

Though both α and β thalassaemia share reduced globin chain synthesis, the pathogenesis of anaemia in severe forms of these conditions is distinct. In severe forms of β thalassaemia (β thalassaemia major), there is severe impairment or absence of β -chain production. As a consequence, excess α -chains accumulate and precipitate in red cell precursors leading to their destruction within the bone marrow. There is also a degree of shortened red cell survival, although ineffective erythropoiesis predominates. Red cells containing fetal haemoglobin (HbF) survive preferentially since there is less globin chain imbalance. The anaemia produces tissue hypoxia, which stimulates erythropoietin production and massive expansion of erythropoiesis in the bone marrow and extramedullary sites.

If untreated, β thalassaemia major leads to severe anaemia, failure to thrive, impaired growth and bony distortion, with death in early childhood. As a result of the expansion in erythropoiesis, there are bone deformities and enlargement of the liver and spleen. Medullary expansion of the facial bones and skull gives rise to a characteristic 'thalassaemic' facies, splaying of the teeth and frontal bossing. With the advent of effective therapy, this picture is now seldom seen in the developed world. The treatment of β thalassaemia major comprises regular blood transfusion to maintain the haemoglobin above 95 g/L. At this level of haemoglobin, bony deformity and hypersplenism due to anaemia are minimized. Leukocyte-depleted blood should be used to prevent alloimmunization against white cell antigens. The blood should be matched for an extended range of blood groups, including ABO, Rh and Kell, to reduce the frequency of alloantibody formation.

An inevitable complication of multiple red cell transfusions is iron accumulation. If untreated, this leads to iron deposition throughout the body, which may cause endocrine dysfunction with diabetes, hypogonadotrophic hypogonadism and hypoparathyroidism, and, ultimately, death due to cardiac failure and liver damage. This may be prevented by the use of iron chelators. Desferrioxamine, which increases both urinary and faecal iron excretion, is administered parenterally, usually by continuous overnight subcutaneous infusion, since it is ineffective orally. Ascorbic acid may enhance urinary iron excretion when given with desferrioxamine. The clinical efficacy of desferrioxamine is limited by its subcutaneous route of administration, and two oral iron chelators are currently licensed for use. Deferiprone is probably less effective than desferrioxamine at removing hepatic iron, but may be particularly effective at removing cardiac iron, and can be used in combination with desferrioxamine. Side-effects include arthropathy and unpredictable neutropenia, and weekly full blood count monitoring is recommended for those taking it. Deferasirox is a newer oral chelator with similar efficacy to desferrioxamine and a good safety profile. The degree of iron overload and the efficacy of

chelation can be monitored by measuring serum ferritin concentration, liver biopsy to quantitate iron and non-invasive measurements using magnetic resonance imaging (MRI) and the superconducting quantum interference device (SQUID). Cardiac and hepatic MRI using T2* and R2 protocols are increasingly used. Accurate quantitation of the volume of transfused blood is also valuable in assessing the degree of iron overload.

With adequate transfusion and chelation therapy, most children with β thalassaemia major attain normal growth and sexual development and survive well into adult life. Allogeneic bone marrow transplantation from human leukocyte antigen (HLA)-matched siblings has been used with considerable success for treatment of β thalassaemia major. The best results, with disease-free survival rates of up to 95%, have been obtained in younger children with little or no evidence of end-organ damage due to iron overload. Trials of gene therapy are taking place in France and the USA, with one patient to date successfully rendered transfusion-independent, although with one cell clone predominating, owing to the inadvertent activation of a potential oncogene.

As in other thalassaemia syndromes, there is considerable heterogeneity in the clinical expression of β globin gene mutations. Much of this heterogeneity is unexplained, but two factors have been identified: first, the extent to which β globin production is impaired, which reflects the nature of the underlying gene mutation, and second, co-inheritance of other interacting genetic determinants that modify clinical severity. Mutations that severely disrupt or abolish β globin synthesis (β^0 thalassaemia) include those that prevent normal splicing of mRNA (splice junction or cryptic splice site mutations) or generate a non-functional mRNA by premature translation termination (nonsense mutations), and small nucleotide deletions or insertions that result in the normal triplet genetic code being read out of frame (frameshift mutations). Other mutations, for example those of the β globin chain promoter, which alter the level of gene expression (transcriptional mutations), usually result in reduced rather than absent β globin synthesis (β^+ or β^{++} thalassaemia). Among the independent genetic factors that ameliorate the clinical effects of β thalassaemia are co-inheritance of α thalassaemia, which partially redresses the imbalance in globin chain synthesis and therefore results in less ineffective erythropoiesis, and hereditary persistence of fetal haemoglobin (HPFH). HPFH can be directly linked to the β globin cluster, and its causes include large deletions and point mutations in the promoter regions of the γ globin genes. Three other major loci which control HbF concentrations have now been identified: the *Xmn1*

BOX 29.1 Causes of β thalassaemia intermedia

Homozygous β thalassaemia

- Homozygous β^{++} thalassaemia
- Compound heterozygote β^+/β^{++} thalassaemia
- Co-inheritance of α thalassaemia
- Co-inheritance of factors promoting increased HbF synthesis

Heterozygous β thalassaemia

- Co-inheritance of triplicated α globin genes ($\alpha\alpha\alpha/\alpha\alpha$ or $\alpha\alpha\alpha/\alpha\alpha\alpha$)
- Dominant β thalassaemia

$\delta\beta$ Thalassaemia

- Homozygous $\delta\beta$ thalassaemia
- Heterozygous $\delta\beta$ thalassaemia/ β thalassaemia

polymorphism in promoter region of the γ globin genes, the *HMP1* locus on chromosome 6q23.3 and *BCL11A* on chromosome 2. These factors account for some cases of the clinical syndrome thalassaemia intermedia, in which there is a variable degree of anaemia, without dependence on regular blood transfusions (Box 29.1).

The heterozygous (carrier) state for β thalassaemia (β thalassaemia trait) is usually harmless. Some of the phenotype characteristics of β thalassaemia heterozygotes are shown in Table 29.2. Most individuals have slightly reduced haemoglobin concentrations and an elevated red cell count. A raised HbA₂ ($\alpha_2\delta_2$) concentration is an important diagnostic marker and distinguishes β thalassaemia trait from α thalassaemia, in which the HbA₂ is normal or low. This presumably reflects a higher output from the δ globin gene, which normally plays a minor role in adult haemoglobin synthesis (normal HbA₂ <3.2%) in the face of reduced β globin production. Similarly, HbF concentrations are frequently slightly raised in β thalassaemia heterozygotes. Rarely, dominant β thalassaemia occurs.

STRUCTURAL HAEMOGLOBIN VARIANTS

Over 700 qualitative variants have been identified, in which the haemoglobin molecule is structurally, and in some cases functionally, altered. These arise from diverse molecular defects, most commonly single amino acid substitutions, but also insertion or deletion of amino acids and polypeptide fusion as a result of recombination between globin genes, for example haemoglobin Lepore. The most common structural variants are haemoglobins S, C, D^{Punjab} and E, which each affect many millions of people worldwide.

TABLE 29.2 The β thalassaemia syndromes

Genotype	Heterozygote			Homozygote		Clinical status
	MCH and MCV ^a	% HbA ₂	% HbF	% HbA	% HbF	
β^0	Reduced	Raised	<4	0	>94	Thalassaemia major
β^+	Reduced	Raised	<4	5–50	50–90	Thalassaemia major
β^{++}	Slightly reduced	Raised or borderline	<2	50–90	10–50	Thalassaemia intermedia

^aMCH, mean cell haemoglobin; MCV, mean cell volume.

Sickle cell anaemia

Sickle cell anaemia (HbSS), caused by homozygous inheritance of structurally abnormal haemoglobin, is one of the commonest genetic disorders with an estimated 60 million carriers worldwide and 230 000 affected infants born annually in Africa alone. The prototypic molecular disease, it was the first human disorder to be understood at a molecular level. The landmark discovery of sickle haemoglobin (HbS) more than half a century ago by Pauling and Itano, was succeeded by demonstration of the underlying single amino acid substitution and the gene mutation responsible. Sickle haemoglobin polymerizes on deoxygenation to produce long molecules that damage and distort the red cell, producing the classic sickle-shaped erythrocytes that give the disease its name. The resulting microvascular occlusion and shortened red cell survival combine to produce a clinical syndrome that comprises anaemia, vaso-occlusion, vasculopathy, haemolysis, inflammation, hypercoagulability, organ damage and susceptibility to infection due to hyposplenism.

A single base change (CAG→CTG) in the sixth codon of the β globin gene determines the substitution of valine for glutamic acid, which results in the structural variant sickle haemoglobin S. While the sickle mutation occurs predominantly in those of African origin, with smaller numbers affected in the countries of the Mediterranean littoral, Saudi Arabia and India, population migration has resulted in its spread globally. The allele frequency of β^s is thought to be high because of protection against the complications of severe malaria in the heterozygous state, as described earlier for thalassaemia mutations.

Although HbSS is the commonest cause of sickle cell disease, this condition occurs with other genotypes, in which there is co-inheritance of a different β globin mutation. The most important of these compound heterozygous states are: HbSC disease, HbS β^+ thalassaemia and HbS β^0 thalassaemia, although about 15 different genotypes have been identified as causing sickle cell disease. Among sickling disorders, HbSS and S β^0 thalassaemia are the more severe clinically, with a greater degree of anaemia (haemoglobin 60–80 g/L) and more severe organ damage. While HbSC and S β^+ thalassaemia are perceived as milder genotypes than HbSS, there is considerable overlap and some types of end-organ damage, particularly proliferative retinopathy, occur more frequently, particularly in HbSC disease. Other genetic factors that may influence disease expression include co-inheritance of α^+ thalassaemia and high fetal haemoglobin (HbF) levels, both of which may ameliorate the severity of sickle cell disease.

The primary pathophysiological event in sickling is intracellular polymerization of deoxy HbS. The $\beta 6$ valine substitution alters the surface charge of haemoglobin, resulting in interaction between haemoglobin tetramers and the formation of 14-stranded polymers. These are then ordered into parallel arrays of fibre bundles, which can be visualized by electron microscopy. The formation of polymer fibres (or gelation) is affected by four main variables: oxygen tension, HbS concentration, temperature and the presence of non-sickling haemoglobins. The mean corpuscular haemoglobin concentration (MCHC)

has a profound effect on the kinetics of sickling through its effect on the delay period between deoxygenation and polymerization. Small increases in haemoglobin concentration, as occur with cellular dehydration, may therefore act as a trigger for the sickling process.

Polymer formation is accompanied by several changes in the membrane of sickle erythrocytes, making them less deformable, increasingly fragile and leaky to cations. These changes are thought to result not only from interactions between sickle haemoglobin and the membrane but also from oxidative damage caused by free radicals, concentrations of which have been shown to be increased in these cells. These membrane changes are initially reversible but become increasingly pronounced as the cell undergoes repeated cycles of sickling and unsickling, culminating in an irreversibly sickled cell (ISC). The percentage of ISCs has been used as a marker for the sickling process but correlates poorly with the clinical condition of the patient. Red cell membrane phospholipids are abnormally distributed in these cells, with the normal inner leaflet amino-phospholipids being rearranged on the outer leaflet of the lipid bilayer. Cation homeostasis is also disordered, with potassium loss exceeding sodium gain, resulting in cellular dehydration and increased concentration of intracellular haemoglobin. This is accompanied by an up to four-fold increase in intracellular calcium. There is increased intravascular haemolysis, leading to a high plasma concentration of free haemoglobin, which binds nitric oxide avidly and results in a functional nitric oxide deficiency; this is implicated in the development of vasculopathy, which contributes to the complications of pulmonary hypertension, priapism and cerebrovascular disease. Other pathological processes include hypercoagulability, inflammation, oxidative stress and reperfusion injury, all of which contribute to the multisystem nature of the disease.

The polymerization of HbS and associated membrane changes cause a marked decrease in the ability of these cells to flow through the microvascular circulation, resulting in compromised oxygen delivery and setting up a vicious circle of worsening vaso-occlusion. There is also evidence of increased adherence of these red cells to vascular endothelium, which would increase deoxygenation and obstruct the flow of other red and white cells through the microvasculature.

The clinical manifestations of sickle cell disease are protean and include both acute and chronic complications. The majority of symptoms are attributable to the effects of vascular occlusion, since the anaemia is usually well tolerated. Sickle cell disease is a multisystem disease. Acute episodes or 'crises' can take several forms, including pain or severe anaemia; acute worsening of anaemia can be due to splenic or hepatic sequestration, or transient erythroid hypoplasia secondary to parvovirus B19 infection. The most common symptom is acute, severe pain, which may be preceded by infection, dehydration and cold exposure, though often no particular trigger is identified. Vaso-occlusion, which commonly involves the bones, results in avascular necrosis of the bone marrow with associated inflammation and increased intramedullary pressure. This results in painful swelling of the hands and feet (dactylitis) in early infancy. In older children and

adults, the juxta-articular areas of long bones, flat bones like the ribs and pelvis and the vertebral column are most commonly affected. The management of acute pain is supportive, ensuring adequate analgesia, hydration and oxygenation. Other acute complications include priapism, retinal artery occlusion and sudden death.

Sickle cell patients are at increased risk of bacterial infection due largely to loss of splenic function and, historically, overwhelming pneumococcal sepsis has been the major cause of early mortality. In developed countries, this risk has been significantly reduced by the introduction of pneumococcal prophylaxis in the form of pneumococcal vaccines and daily oral penicillin. The major cause of mortality after early childhood is the acute chest syndrome, which results from a combination of infection, infarction and fat embolism in the lungs, and is characterized by fever, severe chest pain, dyspnoea and pulmonary infiltrates. This complication needs prompt and vigorous management including oxygenation, hydration, intravenous antibiotics and blood transfusion. Increasing numbers of patients in developed countries are surviving to late middle and old age, and progressive multi-organ failure, particularly involving the kidneys, is an increasingly common form of death. From a very early age, most children with SCD display a number of renal abnormalities, including glomerular hyperfiltration and nephrogenic diabetes insipidus. Many go on to develop significant albuminuria in later childhood, with progressive renal failure contributing to death in about 30% of adults.

Stroke due to occlusion of the large cerebral vessels affects about 11% of patients by the age of 20, with the highest incidence between the ages of two and five years. Without treatment, there is a high rate of recurrence. Long-term transfusion significantly reduces the risk of a further stroke but carries with it the necessity for iron chelation to prevent siderosis. Transcranial Doppler scanning identifies children with early vasculopathy who are at high risk of stroke. Regular blood transfusion of children with such cerebral vasculopathy has been shown to be an effective form of primary stroke prevention, both in clinical trials and practice.

Pulmonary hypertension affects up to 5% of adults with sickle cell disease, and is thought to be associated with an increased risk of premature death. Increased rates of haemolysis have been linked to the pathology, and it may be part of a more general vasculopathy.

Chronic organ damage due to sickling may take several forms. Ischaemic damage to the bones and joints results in progressive destruction which, in the case of avascular necrosis of the hip, may lead to severe disability. Chronic restrictive lung disease may follow recurrent episodes of infection and infarction. As in other chronic haemolytic states, gallstones are common and found in nearly one-third of young adults with sickle cell disease. Proliferative retinopathy can result in bleeding, retinal detachment and blindness. Stasis and occlusion of the small vessels in the lower limbs may cause leg ulceration.

The prognosis for patients with sickle cell disease in developed countries has been transformed by early diagnosis, improved supportive care and, most importantly, prophylaxis against pneumococcal infection. At least 85% of HbSS patients and 95% of HbSC patients in the

USA now survive to 20 years and 50% of patients survive beyond the fifth decade. Relatively little is known about the natural history of the condition in African countries, but the majority of children with sickle cell disease are thought to die before the age of five years. This markedly higher mortality in Africa reflects the importance of environmental factors, and is probably mainly related to malaria, pneumococcal and other infections. Sickle cell disease remains a disease without a cure. Haematopoietic stem cell transplantation has been successfully performed in a few patients but carries a procedure-related risk with 5% mortality. In the absence of reliable predictors of clinical severity, which varies widely among affected patients, this is difficult to justify in most cases. The role of transplantation is likely to increase as the procedure becomes safer and able to draw from a wider range of donors. Human trials of gene therapy are currently planned, but as yet, no patients have been successfully treated.

The search for effective anti-sickling agents has targeted the essential steps in the pathophysiology of sickle cell disease: polymer formation, membrane changes and interactions with the microvasculature. The first approach, aimed at combating polymer formation, either by altering oxygen affinity or increasing HbF concentrations, seems most likely to be successful. Hydroxycarbamide (hydroxyurea), which increases HbF concentrations, has been shown to reduce the frequency of acute pain and acute chest syndrome in randomized controlled trials, and is now an important therapeutic option in adults and children suffering frequent episodes of pain or severe chest problems.

Other structural haemoglobin variants

There are several hundred less common structural haemoglobin variants, most of which are very rare and have no clinical or functional consequences. An exception is the β globin variant haemoglobin E (β^{26} Glu→Lys), which is probably the most prevalent haemoglobin variant, being carried by an estimated 84 million individuals worldwide, mainly in South-East Asia. Haemoglobin E has an electrophoretic mobility similar to that of HbC and HbA₂ at pH 8.9 but can be differentiated by citrate agar electrophoresis at acid pH. Unlike most other structural variants, inheritance of HbE is associated with a β thalassaemia phenotype with microcytosis and hypochromia. This is because the β^E mutation activates a cryptic splice site that inhibits the normal splicing mechanism. Haemoglobin E is also unstable, which may contribute to the unexpectedly severe phenotype that occurs when HbE is co-inherited with β thalassaemia mutations. This compound heterozygous state results in thalassaemia major in about half of cases.

The other common abnormal haemoglobins are C and D^{Punjab}, which affect millions, predominantly in West Africa and the Punjab region of India, respectively. In the homozygous state, there is mild haemolysis, with few if any symptoms. Co-inheritance of both of these variants with HbS results in sickle cell disease.

The unstable haemoglobin variants usually result from neutral substitutions affecting amino acid residues that contact the haem group and generally present as a

congenital Heinz body haemolytic anaemia, for example Hb Köln and Hb Bristol. Heinz bodies are inclusion bodies seen in some red cells consisting of degraded haemoglobin. The diagnosis is made by the heat denaturation or isopropanol precipitation tests and identification of the globin mutation by protein or DNA analysis. Owing partly to the instability of the variant haemoglobin, only half of these variants are detectable by electrophoresis.

Amino acid substitutions involving either α - or β -chains in the vicinity of the haem group can also result in altered oxygen affinity or a propensity to methaemoglobin formation. Haemoglobin variants in which oxygen affinity is significantly increased (e.g. Hb Chesapeake, Hb San Diego) are associated with erythrocytosis (polycythaemia). Low-affinity haemoglobins (e.g. Hb Kansas) and HbM variants (e.g. HbM Boston) cause cyanosis, usually with no associated signs or symptoms of disease. The possible consequences of abnormal haemoglobins are summarized in Table 29.3.

LABORATORY DIAGNOSIS OF HAEMOGLOBINOPATHIES

The primary investigation of a haemoglobinopathy should include a full blood count, peripheral blood film and haemoglobin electrophoresis (see Fig. 29.4). The full blood count allows assessment of haemoglobin formation, as judged by the red cell indices, the mean cell volume (MCV) and mean cell haemoglobin (MCH). Microcytosis (MCV <76 fL) and hypochromia (MCH <27 pg), in the face of a normal or raised red cell count ($>5.5 \times 10^{12}/L$) in an iron-replete patient, suggest a diagnosis of thalassaemia. In the case of β thalassaemia major or intermedia, this is associated with a significant degree of anaemia, whereas in thalassaemia trait the haemoglobin is usually normal or only marginally reduced. Examination of a blood film stained by a Giemsa method may be helpful in confirming the diagnosis. However, definitive diagnosis often rests on electrophoretic or chromatographic analysis of haemoglobin in red cell haemolysates. Cellulose acetate electrophoresis at alkaline pH (8.9–9.1) is the most widely used method, being simple, rapid, inexpensive and effective in separating the common haemoglobin variants. In homozygous sickle cell anaemia, HbS predominates. A variable amount of HbF is present, higher proportions

(>10%) generally being associated with a milder clinical course. Haemoglobin A₂ concentration is usually normal.

Solubility testing based on the reduced solubility of deoxy-HbS in the presence of reducing agents, for example sodium dithionite, has a limited role in the diagnosis of sickle cell disease since it does not differentiate the homozygous disease and carrier states. In an emergency situation, a positive solubility test taken in conjunction with a significantly reduced haemoglobin and typical red cell morphology points strongly towards a diagnosis of sickle cell disease. This should be confirmed by haemoglobin analysis at the earliest opportunity. Several other variants, including HbD, HbG and Hb Lepore, have an electrophoretic mobility identical to that of HbS on cellulose acetate but may be distinguished by the negative sickle solubility test and citrate agar gel electrophoresis at acid pH (6.0). Similarly, haemoglobins C, E and O, which co-migrate on cellulose acetate at alkaline pH, can be differentiated by citrate agar electrophoresis. Both HbE and Hb Lepore are associated with thalassaemic red cell indices, which further aids their distinction from electrophoretically similar variants. Isoelectric focusing improves the resolution of some structural variants and can also be used for neonatal screening of eluates from Guthrie (dried blood spot) cards, since it reduces interference from methaemoglobin present in such samples.

High performance liquid chromatography (HPLC) is a fast and sensitive method for separation and quantitation of haemoglobins that, in some cases, allows identification of variants not possible by other techniques. Since it is largely automated and requires minute quantities of sample, HPLC has become the method of choice for large-scale population testing such as neonatal screening programmes. Universal neonatal screening for sickle cell disease has been used in the USA and England for some time, and similar programmes are being introduced into other European, Middle Eastern and African countries, depending on the prevalence of the conditions and the resources available. Such programmes have resulted in significant benefits in terms of reduced mortality and morbidity due to improved care, early implementation of prophylaxis against pneumococcal infection and parental education. In β thalassaemia, the proportion of individual haemoglobins varies with the underlying genotype. Homozygous β^0 thalassaemia is associated with

TABLE 29.3 The functional and clinical consequences of abnormal haemoglobins

Functional abnormality	Clinically expressed ^a	Clinical disorder	Example	Molecular abnormality
None	No	None	HbG Philadelphia	$\alpha 68$ Asn→Lys
Polymerization with reduced solubility	Hom/Co het	Sickle cell disease	HbS	$\beta 6$ Glu→Val
Unstable	Het	Haemolytic anaemia	Hb Köln	$\beta 98$ Val→Met
Increased O ₂ affinity	Het	Erythrocytosis	Hb Chesapeake	$\alpha 92$ Arg→Leu
Decreased O ₂ affinity	Het	Cyanosis	Hb Kansas	$\beta 102$ Asn→Thr
Methaemoglobinemia	Het	Blue discolouration	HbM Saskatoon	$\beta 63$ His→Tyr
α Thalassaemia	Hom/Co het	Thalassaemia intermedia	Hb Constant Spring	$\alpha + 31C^b$ (142 Stop→Gln)
β Thalassaemia	Co het	Thalassaemia intermedia/major	Hb Lepore-Boston	δ (1–87) β (116–146) fusion

^aThe abnormality may be expressed in the homozygous (Hom), heterozygous (Het) or compound heterozygous (Co het) state.

^bThis chain terminator mutation results in an α globin chain with an additional 31 amino acids.

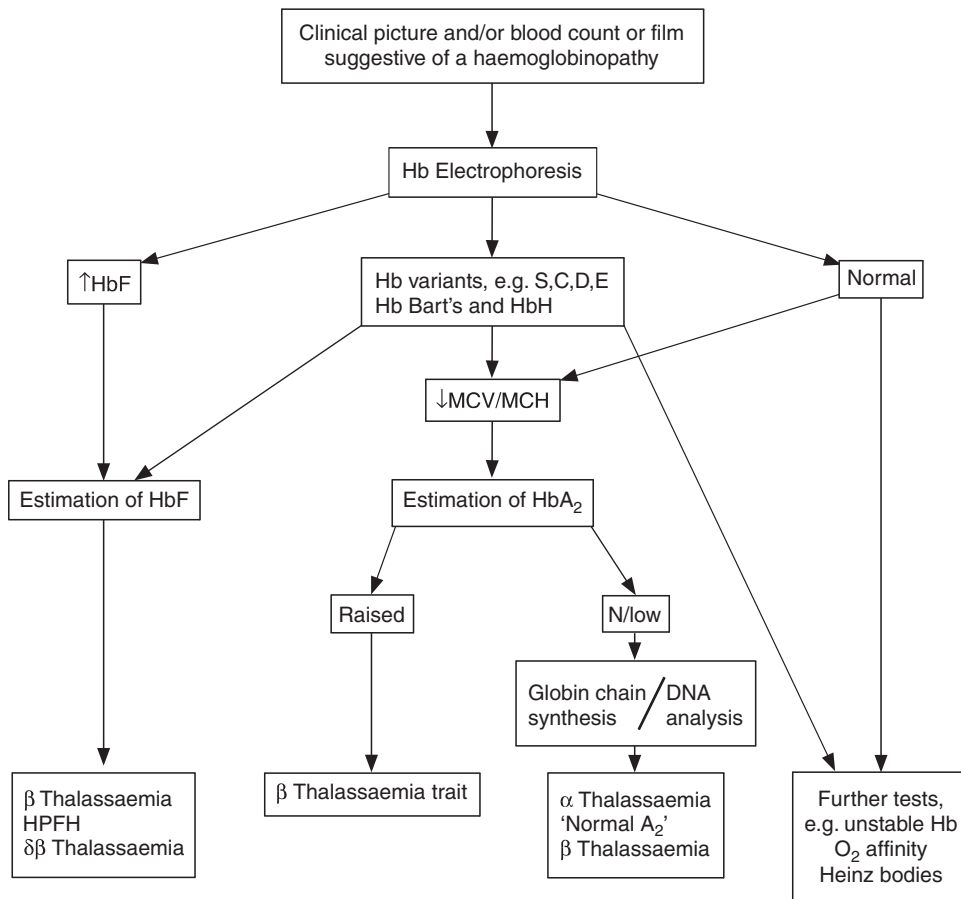


FIGURE 29.4 ■ A simple algorithm for the diagnosis of haemoglobinopathies.

a predominance of HbF, absence of HbA and variable amounts of HbA₂ (range 1.0–6.0%, mean 1.7%). In individuals with homozygous β⁺ thalassaemia or compound heterozygous β⁰/β⁺ thalassaemia, a variable amount of HbA is present. Haemoglobin F is increased and distributed heterogeneously among red cells.

Accurate quantitation of HbA₂ by HPLC or microcolumn chromatography is essential for the diagnosis of β thalassaemia trait, in which the HbA₂ is elevated, typically >3.5%. Carriers of 'normal A₂' or 'silent' β thalassaemia due to mild β gene defects or co-inheritance of a δ gene mutation in *cis* or *trans* cannot easily be distinguished from α thalassaemia by conventional screening methods and require investigation by specialized techniques, including in vitro globin chain synthesis and DNA analysis. Analysis of globin chain synthetic ratios by tritiated leucine incorporation and carboxymethylcellulose chromatography is the definitive way of identifying individuals with thalassaemia, although it is rarely used because of its laborious nature.

The α thalassaemias are characterized electrophoretically by the presence of the fast moving variants, Hb Bart's (γ4) and HbH (β4), which are most obvious in neonatal samples. In hydrops fetalis owing to homozygous α⁰ thalassaemia, Hb Bart's predominates and is found in smaller amounts in other α thalassaemia syndromes in the neonatal period. Haemoglobin H may also be detected by staining for HbH inclusion bodies. The diagnosis of clinically silent forms of α thalassaemia is often one of

exclusion, being made on the basis of the subject's ethnic origin, microcytic hypochromic red cell indices and a normal or low HbA₂ concentration in the presence of normal iron status. Definitive diagnosis can be made by DNA analysis, which can also often distinguish α⁰ and α⁺ thalassaemia.

While the majority of haemoglobinopathies can be diagnosed by haemoglobin electrophoresis, variants caused by amino acid substitutions that do not alter charge, such as those found in some unstable haemoglobins or haemoglobins with altered oxygen affinity, may escape detection. Further investigations that may be helpful in this context include assessment of haemoglobin instability and oxygen affinity. High throughput DNA analysis has made this a feasible way of screening for globin mutations in richer countries, with techniques such as multiplex ligation-dependent probe amplification allowing identification of large gene mutations which were previously only detectable using Southern blotting. Haemoglobin mass spectrometry can also be used to identify abnormal globins by measuring their mass accurately, with particular potential use in screening programmes which already use mass spectrometry.

The identification of couples at risk for major haemoglobinopathies by antenatal or preconceptional screening permits informed reproductive choice with the option of prenatal diagnosis. In most cases, this can now be accomplished in the first trimester by detection of mutant globin genes in chorionic villous DNA. In several countries,

notably Cyprus where the carrier rate for β thalassaemia reaches 12%, this has led to a marked decline in the birth incidence of haemoglobin disorders. Preimplantation genetic diagnosis is increasingly used to allow selection of unaffected embryos, although it continues to be a demanding and expensive process that is not applicable to most couples. Attempts continue to develop non-invasive prenatal diagnosis using fetal cells and DNA in maternal blood, although this is currently not technically feasible as a routine procedure for the haemoglobinopathies.

CONCLUSION

Haemoglobin is arguably the best studied of all human proteins. It is a tetramer of two pairs of globin chains. Each chain binds a molecule of haem, to which a molecule of oxygen can become reversibly bound. The affinity of haemoglobin for oxygen can be modified by various physiological factors in ways that respond to the requirements of tissues for oxygen under different circumstances.

Inherited disorders of haemoglobin synthesis can be qualitative or quantitative. More than 400 structural variants have been described, of which the most important is haemoglobin S, the haemoglobin of sickle cell disease. Quantitative abnormalities of globin chain synthesis cause the thalassaemias. Advances in diagnosis and treatment have dramatically altered the prognosis for many affected patients, although there is still a limited range of treatments available.

Further reading

Dacie J. The hereditary haemolytic anaemias. Part 2. 3rd ed. The haemolytic anaemias. vol. 2. Edinburgh: Churchill Livingstone; 1988.

Rees DC, Williams TN, Gladwin M. Sickle-cell disease. *Lancet* 2010;376:2018–31.

Serjeant GR, Serjeant BE. Sickle cell disease. 3rd ed. Oxford: Oxford University Press; 2001.

Weatherall DJ, Clegg JB. The thalassaemia syndromes. 4th ed. Oxford: Blackwell Science; 2001.

Each of these texts provides a comprehensive account of the pathology, diagnosis and management of the haemoglobinopathies.

Immunology for clinical biochemists

Joanna Sheldon • Rachel D. Wheeler • Pamela G. Riches

CHAPTER OUTLINE

THE IMMUNE SYSTEM 560

Introduction 560

Immune responses 560

The innate immune system 561

The adaptive immune system 562

Complement 570

Acute phase proteins 571

Cytokines 572

Conclusion 575

DISEASES OF THE IMMUNE SYSTEM 575

Introduction 575

Immune deficiency 575

Allergies 580

Autoimmune diseases 582

LYMPHOID MALIGNANCIES 588

B lymphocytes and monoclonal proteins 588

β_2 -Microglobulin 594

B cell malignancies 594

INFECTION AND SEPSIS 597

Diagnosis and monitoring of infections 598

TRANSPLANTATION 599

Organ transplantation 599

Stem cell transplantation 600

CONCLUSION 600

APPENDIX 602

THE IMMUNE SYSTEM

Introduction

The human immune system monitors the internal and external environment at a molecular level and relays information within itself and to other systems regarding the molecular nature of the environment. Within this process, the system has evolved so that certain stimuli elicit a reaction (an immune response) while others are tolerated, eliciting no response. Microbes that invade our bodies can replicate every 20 min, giving them enormous potential to evolve more virulence. Humans reproduce approximately every 20 years with considerably less potential for population adaptation. However, an immune system that uses overlapping and interwoven processes, and can evolve by somatic mutation, allows an individual to adapt to a changing microbial environment. The immune response, like any other biological process, attempts to restore homeostasis but may itself, if inappropriate or persistent, result in tissue damage and disease.

Over the last 20 years, there has been a rapid development of our knowledge of immunology, particularly at the biochemical and molecular level, in terms of the sophisticated intercellular interactions that accompany any immune response. This chapter will make no attempt to summarize this vast body of knowledge but rather aims to convey broad concepts in describing immune responses, the immune system that makes the responses and the relevance of these to clinical laboratory medicine. The spectrum of immunological diseases includes genetic

deficiencies, autoimmune diseases, allergy, malignancy, inflammation and infection. Manipulation of immune responses by inducing immune activation or suppression has great importance in many areas, including immunization and transplantation.

Immune responses

There are two major types of immune response: innate and adaptive. The innate immune response shares many components with the adaptive response but differs fundamentally in that no specific antigen recognition is required. Instead, it involves recognition of molecular sequences that are widely expressed in nature, for example on microbial cells (lipopolysaccharide) and on damaged tissue (heat shock proteins).

The adaptive response is antigen driven and requires prior antigen exposure and recognition of the antigen by T and B lymphocytes through cell surface receptors. On first contact with any antigen, the system reacts relatively slowly and transient infection may occur. Immunologically, this process is termed *immunization*. The initial response is known as the *primary immune response*: it also involves acquisition of memory of the exposure. As a result, on second contact with an antigen there is a more rapid response (the *secondary immune response*) and there is no infection: the host has acquired *immunity*. These responses involve a combination of cells, tissues and soluble mediators that are described more fully below. In order to understand immune responses, it is first necessary to consider two important terms: antigen and clonality.

Antigens

Most biological materials, including proteins, polysaccharides, glycoproteins and polypeptides can be antigens. How the antigen influences the immune system is related to its reaction with lymphocytes, which will determine whether it stimulates an immune response (i.e. it is an immunogen) or is tolerated (i.e. it is a tolerogen). For example, human albumin is tolerated by the human immune system but on injection into a rabbit, it will stimulate an immune response such that the rabbit produces antibodies to the protein. These antibodies, which react with human albumin, but not, for example, with horse albumin or with the rabbit's own albumin, are described as being specific for human albumin.

The antibodies produced are not directed against the entire molecule but to short, overlapping amino acid sequences termed epitopes. A large protein molecule such as albumin will have many hundreds of epitopes and, generally, the larger the protein, the more immunogenic it is.

Clonality

Both T and B lymphocytes have the property of antigen recognition through molecules expressed on their cell surfaces (antigen receptors). This is immunoglobulin (Ig) on B cells and the T cell receptor (TCR) on T cells. On any one cell, all the antigen receptors will be identical. Generally, this unique receptor is able to recognize one or a few epitopes. It is estimated that approximately 10^{10} different antigenic specificities may potentially be encountered during the lifetime of an individual, requiring approximately this number of unique receptors. The development of this very large number of lymphocyte receptors occurs before antigen contact by random rearrangements of the immunoglobulin and TCR genes. The molecular process by which this is achieved will be described in more detail in a later section. Immune responses are driven by the presence of antigen. Engagement of a compatible receptor with antigen provides signals for further proliferation and differentiation. The affinity of the antigen-receptor binding is crucial in determining the nature of that immune response. This is the molecular basis for the establishment of immune responses and immunological memory.

Additionally, in the B cell, the DNA that codes for the part of the Ig receptor that binds antigen is hypermutable, so that during proliferation, somatic mutations within the receptor DNA sequences occur frequently. These mutations may result in a receptor that has a higher affinity for the antigen and, as a result, the clone bearing that receptor will compete more successfully for the antigen and receive a stronger signal, so that it is preferentially selected and responds more quickly. The secreted immunoglobulin from the B cell will similarly show this higher affinity. This process is termed affinity maturation. Immunity to many potentially lethal or harmful human diseases, for example measles, polio and smallpox, is mediated by such high-affinity immunoglobulin (antibody).

The innate immune system

One of the major roles of the immune system is to neutralize and destroy invading pathogens. These pathogens – bacteria, viruses, fungi, protozoans and worms – vary in their size, typical route of entry into the body and mechanism for evading the immune system. Generally, they enter the body across the mucosal surfaces of the respiratory system and gut and across the skin; it is in these areas of the body where the innate immune system forms the first-line of defence from invading pathogens. The major components of the innate immune system are:

- mechanical barriers, e.g. intact skin, mucous membranes
- sebaceous secretions; contents include fatty and other acids (hence skin acidity)
- lysozyme (in tears)
- gut acidity
- urine acidity
- intestinal motility
- IGA (in tears, at mucosal surfaces)
- mucus
- ciliated surfaces
- normal bacterial flora
- acute phase proteins
- interferons
- other molecules (including defensins, lactoferrin, peroxidases)
- neutrophils and complement, although these overlap with the acquired immune system.

Intact mucosal surfaces give a high degree of protection against pathogens, and if breached, for example by surgery, burns or mechanical damage, significant infection can occur. The skin is relatively dry with a high salt content; sweat and sebaceous secretions contain fatty acids, triglycerides, lactic acid, amino acids and ammonia that have antimicrobial activity. Mucus and the cilia of the respiratory tract act together to trap and sweep inhaled particles towards the mouth and nose, where they can leave the body or where they can be swallowed to be inactivated by the gastric acid. The gastrointestinal tract has a combination of innate defence mechanisms such as the alkalinity of saliva and the strongly acidic conditions and enzymic activity of the stomach. Mucopolysaccharides in the secretions of the genital tract help make it inhospitable to pathogens. The eyes are bathed in tears containing the enzyme lysozyme, and the rate of tear production can increase to flush away any potential invading pathogens. Similarly, the urinary tract has a high flow rate that helps to keep it free from pathogens.

Many areas of the body are colonized with commensal organisms. It is estimated that a normal gut is colonized with 10^{14} bacteria, which form the normal (commensal) bacterial flora and produce chemicals that are antimicrobial and also compete with potentially pathogenetic organisms for essential nutrients. Commensal organisms may cause infection – usually called opportunistic infections – particularly if they get into areas where they do not normally live or if the immune system is compromised.

The innate immune system overlaps with the adaptive immune system (Fig. 30.1), with both systems using the complement cascade, cytokines, phagocytic cells, natural

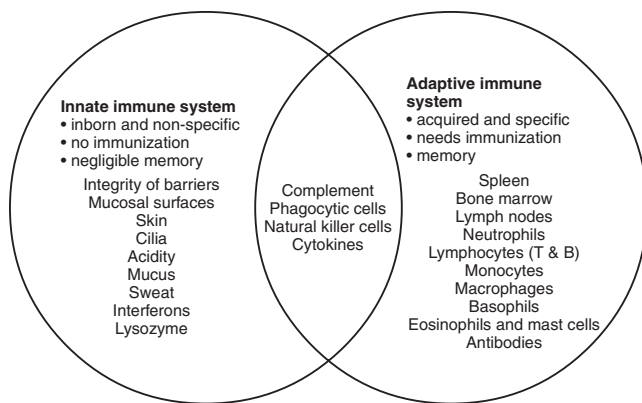


FIGURE 30.1 ■ The overlap between the innate and adaptive immune systems.

killer cells and acute phase proteins. These components can act alone in an innate response to destroy invading pathogens, but the mechanisms are non-specific and slow and are significantly enhanced by involvement of the adaptive immune system.

The adaptive immune system

An adaptive immune response involves an integrated system, in which there are three elements.

1. **Lymphoid tissue**, wherein the components of the immune system are made and where immunological processing occurs.
2. **Cells with specific functions**, including specific antigen recognition and presentation.
3. **Soluble mediators** that participate in immune reactions and can interact with other systems of the body.

The components of the adaptive immune system are summarized in [Table 30.1](#).

TABLE 30.1 The components of the adaptive immune system

Lymphoid tissue	Cells	Soluble mediators
Embryonic^a	Neutrophils	Antibodies
Yolk sac	Monocytes	Complement proteins
Fetal liver	Macrophages	Cytokines
	T lymphocytes	Acute phase proteins
Primary	Natural killer cells	Inflammatory mediators:
Bone marrow		Tryptase
Thymus	B lymphocytes	Histamine
	Basophils	Platelet activating factor
Secondary	Eosinophils	Prostaglandins
Spleen	Mast cells	Leukotrienes
Lymph nodes and lymphatics	Dendritic cells	Soluble receptor molecules
Tonsils	Cell bound recognition molecules	
Adenoids		
Lymphoid tissue in the skin and mucosa of the bronchus, gut and urogenital tract	(HLA or MHC) Cell receptor molecules	

^aAfter the early fetal development, all the cellular elements of the blood arise from haematopoietic cells in the bone marrow. HLA, human leukocyte antigen; MHC, major histocompatibility antigen.

Lymphoid tissue

Lymphoid tissue is distributed around the body ([Fig. 30.2](#)), with the blood and the lymphatics forming the lymphoid circulatory system. The bone marrow and the thymus are primary lymphoid organs, the bone marrow being the major site of production of all the cells of the immune system. The bone marrow is also the site of maturation of B lymphocytes. Immature T lymphocytes leave the bone marrow to develop further in the thymus (a bilobed organ in the upper anterior mediastinum that diminishes in size with age). The secondary lymphoid organs are found throughout the body, but particularly at junctions of the lymphatics, for example in the groin and axillae. This arrangement enables specialized antigen-presenting cells (APCs) to show antigens to an enormous number of lymphoid cells as they ‘traffic’ around the immune system. Secondary lymphoid tissue can be diffuse aggregates of cells or can be more organized into identifiable areas, for example the tonsils and Peyer patches in the gut. Lymph nodes also have an identifiable structure, with B cells residing mainly in the outer cortex and T cells mainly in the paracortex. Upon antigen stimulation, primary follicles within lymph nodes develop into secondary follicles with germinal centres. Activation of the immune system causes the cells within the lymph nodes to differentiate and proliferate, resulting in enlargement of the lymph nodes (lymphadenopathy). This is sought on physical examination of patients as a sign of increased activity of the immune system, for example in infection and malignancy. A large proportion of lymphoid cells are found in the gut and respiratory systems (gut and mucosal associated lymphoid tissue: GALT and MALT), where the antigenic burden is high.

Although specific immune responses are made within lymphoid tissues, other non-lymphoid organs have a role to play; for example, the liver is the major site of production of the complement proteins and acute phase proteins.

Cells

All cells of the body interact with the immune system. For example, any damaged nucleated cell has the capacity to make cytokines, and red blood cells have complement and antibody receptors on their surface that, owing to their enormous numbers, represent an important removal mechanism for immune complexes. Immune cells circulating in the blood are represented by the white cell population. Normally, the total white cell count of an adult is $4\text{--}11 \times 10^9/\text{L}$. This total is made up of a number of different cell types as shown in [Table 30.2](#); the differential white cell count represents the number and proportion of each of these types. Each cell type has particular functions within the immune system: the major functions are listed in [Table 30.2](#).

Each cell type has a characteristic morphology that enables it to be identified in a stained blood film examined under a microscope. [Figure 30.3](#) shows diagrams of the various types of white blood cell with their major progenitor cell types. The cell membranes are covered with molecules that enable the cells to interact

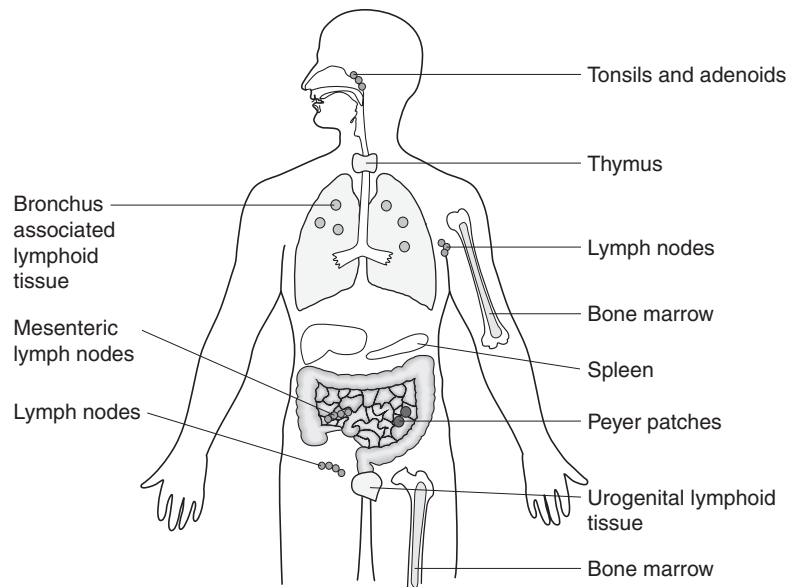


FIGURE 30.2 ■ The lymphoid organs.

TABLE 30.2 The major cells types, their approximate percentage of the differential white cell count and their major functions

Cell type	Reference range ($\times 10^9/L$)	Function
Total white cell count	4–11	
Neutrophil	1.6–7.5	Attracted to site of inflammation by chemotactic factors Most important cell type for phagocytosis and killing of bacteria Opsonization with antibody and complement enhances phagocytosis
Lymphocyte	1.5–4.0	Key role in orchestrating the immune response and producing soluble mediators of immunity
T cell (CD3+)	0.8–2.66	T helper cells (CD4+) control the development of an immune response by secreting cytokines Cytotoxic T cells (CD8+) can directly kill cells, e.g. that are infected with virus Regulatory T cells modify immune responses
B cell (CD19+)	0.1–0.6	Antibody production
Natural killer cell (CD56+, CD16+)	0.05–0.60	Can directly kill cells, e.g. infected with virus, tumour cells
Monocyte	0.1–0.8	Important phagocytic cells and can act as antigen presenting cells Cytokine production – particularly in inflammation
Basophil	<0.1	Major role is in allergic responses via release of mediators, e.g. histamine May be important in defending against parasites
Eosinophil	0.04–0.45	Some phagocytic function Major role in hypersensitivity reactions via release of variety of mediators Activation of mast cells

with their environment. They may be constitutively expressed (i.e. expressed in the unstimulated state) and this expression be increased on activation, or they may be induced when a cell is activated. Cells have many thousands of molecules of a whole variety of cell receptors on their surfaces. The cell surface molecules are classified by their cluster of differentiation (CD) numbers, with a different CD number for individual types of cell membrane protein. (The designation + after a CD number indicates that a specific antigen is expressed by that cell. For example, all T cells are CD3+; T helper cells are CD4+.) The cellular expression of these markers is used to provide a more detailed picture

of cell populations, for example whether they are immature or mature, whether activated, how they are functioning and whether they are normal or malignant. Aberrant expression of the CD markers is used to classify lymphoid malignancies while reduced expression of CD antigens can be used in the investigation of immunodeficiency, for example, CD4 counts to monitor HIV infection. The technique (known as immunophenotyping), uses specific antibodies to the CD markers and flow cytometry to count and quantify cell populations. There are approximately 400 different human proteins that have been assigned a CD number. The most important immunological markers, their major functions

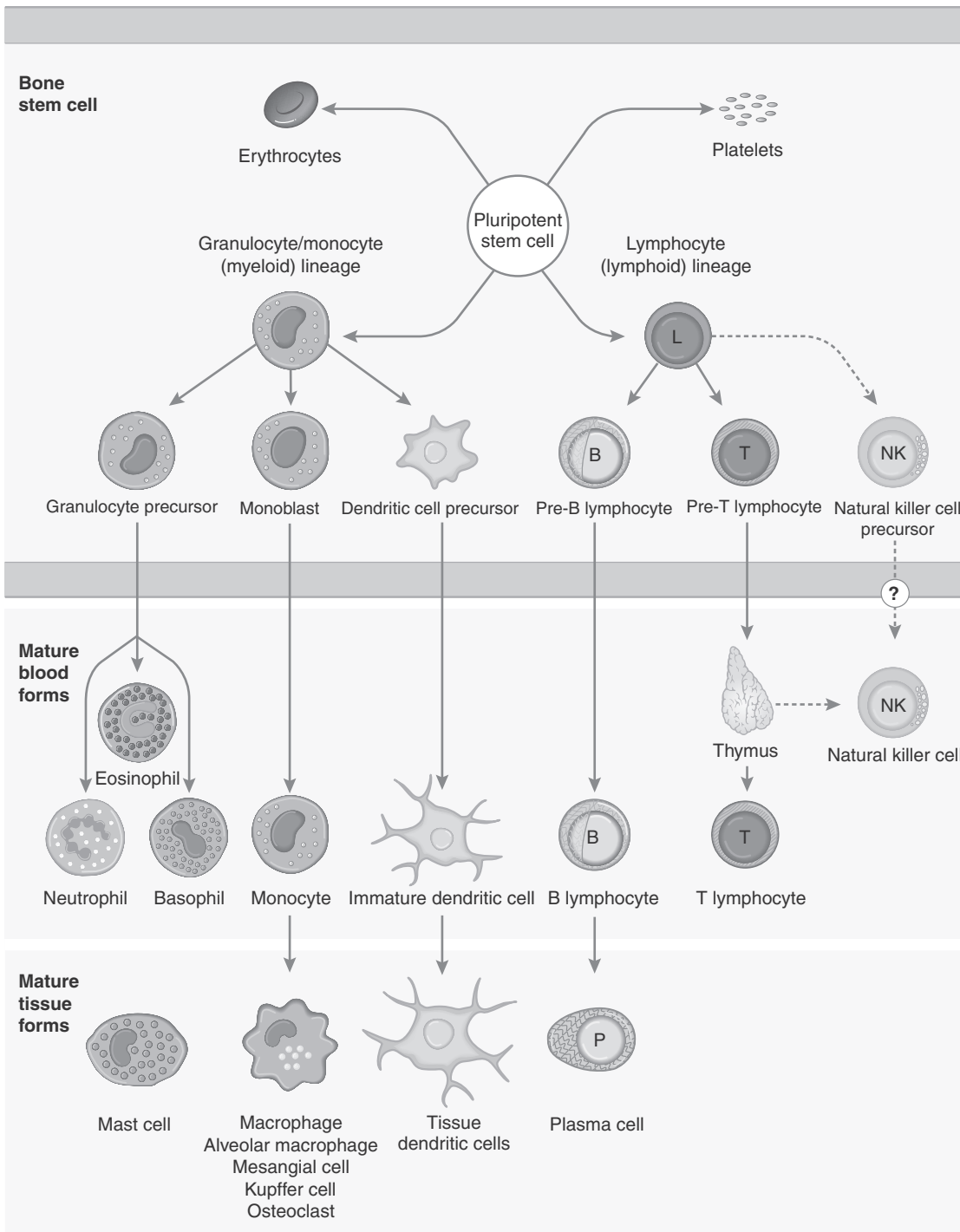


FIGURE 30.3 ■ Major types of white blood cell and their generation.

and the cells types on which they are predominantly expressed are shown in [Table 30.3](#).

The cells of the immune system arise from haematopoietic stem cells in the bone marrow. Individual stem cells respond to a number of positive or negative feedback mechanisms, for example colony stimulating factors, and are driven to proliferate and differentiate along committed pathways to give rise to populations of lymphocytes, granulocytes, APCs etc. The major processes are outlined in [Figure 30.3](#).

Neutrophils. Neutrophils, in common with basophils and eosinophils, can be shown to contain a large number of cytoplasmic granules when stained and examined by light microscopy and are thus termed granulocytes. They have a multilobed nucleus (polymorphonuclear or PMN, hence their alternative name, polymorphs) and are the most abundant white cells in the circulation. They show high expression of adhesion molecules that enable them to roll along blood vessel walls. If expression of adhesion molecules on the vascular surface is upregulated,

TABLE 30.3 The major CD phenotypes, their normal functions and cell types where they are normally expressed

Phenotype	Function of protein	Cells where expressed
CD3	Part of the T cell receptor	All T cells (pan T cell marker)
CD4	Helps T cells to hold on to antigen-presenting cells (HLA class II)	T helper cells
CD8	Helps T cells to hold on to antigen-presenting cells (HLA class I)	Cytotoxic T cells
CD19	B cell co-receptor	B cells
CD20	B cell activation and differentiation	B cells
CD45	Cell differentiation, signal transduction and lymphocyte activation	All white blood cells (leukocyte common antigen)
CD16	Binds $F_c \gamma$ RIII (binds IgG F_c)	Natural killer cells, neutrophils, monocytes
CD32	Binds $F_c \gamma$ RII (binds IgG F_c)	Neutrophils, macrophages, B lymphocytes, eosinophils
CD64	Binds $F_c \gamma$ RI (binds IgG F_c)	Monocytes, macrophages, activated neutrophils
CD89	Binds $F_c \alpha$ R (binds IgA F_c)	Monocytes, macrophages, neutrophils
CD21	CR2 (binds C3d)	B cells (some)
CD35	CR1 (binds C3b, C4b)	Neutrophils, monocytes, B cells, erythrocytes
CD25	IL-2R α chain	Activated T and B cells
CD56		Natural killer cells
CD40	Binds to CD40 ligand (CD154); B cell proliferation and class switching	B cells, antigen-presenting cells
CD11a (+CD418)	Cell adhesion (to other cells)	Leukocytes
CD11b (+CD18)	Cell adhesion (to other cells) CR3	Monocytes, neutrophils, natural killer cells
CD11c (+CD18)	Cell adhesion (to other cells) CR4	Monocytes, neutrophils, natural killer cells, some T and B cells
CD18 (+ CD11 a, b and c)	Cell adhesion (to other cells)	Leukocytes, monocytes, neutrophils, natural killer cells, some T and B cells
CD62E	E selectin cell adhesion (to other cells)	Endothelium, platelets
CD62L	L selectin cell adhesion (to other cells)	Neutrophils
CD62P	P selectin cell adhesion (to other cells)	Endothelium
	B cell proliferation and class switching	Activated T cells

IL, interleukin; HLA, human lymphocyte antigens.

for example because of tissue damage or infection, neutrophils are attracted to the site of inflammation (by chemotactic factors) and leave the blood stream and enter the tissues to engulf the invading pathogens or damaged tissue components, and destroy them, in a process termed phagocytosis.

The stages in phagocytosis are as follows.

1. **Cells move to a site of injury** or inflammation, attracted by chemicals released during the inflammatory process. Adhesion molecules, for example on the endothelium, are upregulated to hold the phagocytic cells and prevent them from moving away.
2. **At the site of injury**, the particle or microorganism, coated with antibody and complement, adheres to the phagocyte via C3b and $F_c \gamma$ receptors.
3. **The adherence of the particle to the phagocyte** results in the activation of the phagocyte's membrane, initiating a change in its shape and leading to internalization of the foreign material (phagocytosis: literally, 'eating' the particle).
4. **As the particle is taken into the cell**, it becomes surrounded with a membrane, forming a phagocytic vesicle or phagosome.
5. **The phagosome fuses with lysosomes** to form a phagolysosome, in which the microorganism is killed, and it, or other ingested material, is digested.

6. **Completely digested material** is reused by the cell; incompletely digested material is released (sometimes accompanied by potentially harmful enzymes).

If phagocytic cells are incubated in vitro with, for example, bacteria, phagocytosis will occur, but the process is significantly enhanced if the bacteria are coated with complement or antibody. This coating is called opsonization and, in simple terms, it enables phagocytic cells to adhere to the marked particles (giving the process specificity) and enables them to change shape to engulf the particles. Many cell types are capable of phagocytosis, the 'professional' phagocytic cells being neutrophils and cells of the monocyte/macrophage lineage. Many macrophages home to particular tissues where they develop highly specialized functions, e.g. Kupffer cells (in the liver), mesangial cells (kidneys), microglia (brain) and Langerhans cells (epidermis).

Basophils and eosinophils. These two cell types are also granulocytes and are particularly important in defending against multicellular pathogens, for example parasitic worms. This defence mechanism relies on the mediators in the cytoplasmic granules rather than the phagocytic functions of the cells. A close relation of the basophil, the mast cell, is tissue-based and important in allergic responses.

Monocytes. Monocytes are immature macrophages that are travelling from their site of production in the bone marrow through the bloodstream to the tissues where they will become fixed macrophages. Potentially, they are phagocytic cells but they do not seem to act as phagocytes in the bloodstream. They have receptors that recognize many bacterial products so are very effective cells in innate immunity. Engagement of these receptors stimulates release of proinflammatory cytokines such as interleukin 1 (IL-1), IL-6 and tumour necrosis factors (TNFs). These cytokines are able to drive many aspects of inflammation and specific immunity. Monocytes therefore provide a valuable second-line defence mechanism once infectious agents have breached the mucosal defences and entered the bloodstream.

Lymphocytes. The major subpopulations of lymphocytes are B and T cells. A haemopoietic stem cell first gives rise to a lymphopoietic stem cell, which in turn becomes committed to the B or T cell lineage. Lymphocytes are further divided into subsets on the basis of the presence of particular cell surface molecules (see Table 30.3). B cells (CD19+) can differentiate into antibody (immunoglobulin)-producing plasma cells in response to exposure to antigen. T cells leave the bone marrow to mature in the thymus. They undergo further differentiation to become either T helper cells (CD4+) that ultimately control the immune response, through cytokine production or cytotoxic T cells (CD8+) that have an important role in killing viruses, fungi and infected cells. Other T cell types are becoming increasingly well recognized as having important immunological roles; for example, that of regulatory T cells (T_{reg}) in the modulation of autoreactive T cells, and of Th17 (T helper) cells in inflammatory responses.

There are two main types of T helper cells (Th1 and Th2), which are defined by the cytokines they produce. Th1 cells predominantly produce IL-2, TNF α/β and γ -interferon, and drive cell-mediated immunity by cytotoxic T cells, macrophage stimulation and inflammation, and production of immunoglobulins, such as IgG1 and IgG3 by B cells. Th2 cells predominantly produce IL-4, 5, 6, 10 and 13, and are important in driving IgE responses and the growth of mast cells and eosinophils.

Natural killer (NK) cells (CD16+ and CD56+) are cytolytic cells of the lymphocyte lineage, acting in a slightly different way from the cytotoxic Tc cells.

B and T lymphocytes are at the heart of the adaptive immune response by virtue of their antigen receptors (Ig and T cell receptor, TCR). A comparison of the B and T cell receptors is shown in Table 30.4. The property of antigen binding is also shared by the human leukocyte antigens (HLA). These three sets of antigen-binding molecules are all products, or partial products, of the immunoglobulin gene superfamily; they are all composed of one or more homologous domains of approximately 110 amino acids, as illustrated in Figure 30.4.

Antigen recognition

Immunoglobulins. Immunoglobulins are proteins produced by cells of the B-lymphocyte lineage, either in a membrane-bound form as an antigen receptor, or as a secreted product that shows antibody activity. The membrane-bound and secreted forms differ only in the terminal amino acid residues that serve to anchor the protein into the membrane or allow secretion. The basic monomeric immunoglobulin unit consists of two identical heavy chains and two identical light chains arranged in a Y-shape (see Fig. 30.5). In humans, there are five possible heavy chains (γ , α , μ , δ and ϵ) that give the five immunoglobulin classes (IgG, IgA, IgM, IgD and IgE). In addition, there are four subclasses of IgG and two subclasses of IgA. There are two light chain types (κ and λ).

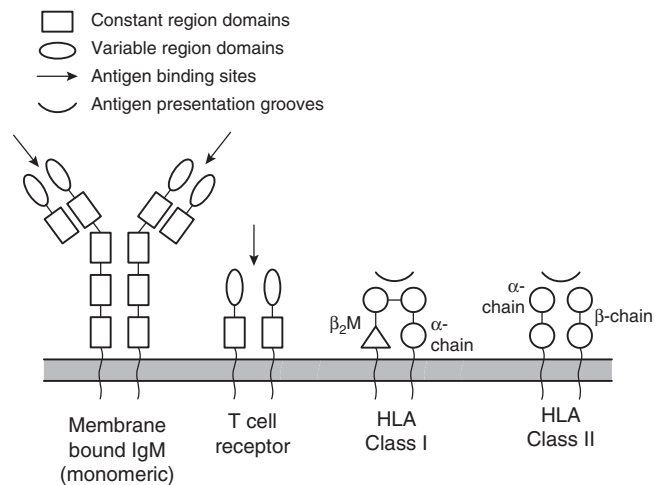


FIGURE 30.4 ■ The antigen presenting and binding molecules. β_2M , β_2 -microglobulin; HLA, human leukocyte antigen.

TABLE 30.4 Comparison of B cell receptor and T cell receptor

	B cell receptor (membrane-bound IgM)	T cell receptor
Similarities	Variation created by gene rearrangement Part of a protein complex including accessory chains and signalling adaptors Variable and constant domains	
Differences	Two heavy chains and two light chains Somatic hypermutation allows affinity maturation Class switching from IgM to other immunoglobulin classes Binds directly to specific antigen	Two polypeptide chains ($\alpha\beta$ or $\gamma\delta$) No somatic hypermutation No class switching Binds to specific antigen only when bound by appropriate HLA molecule

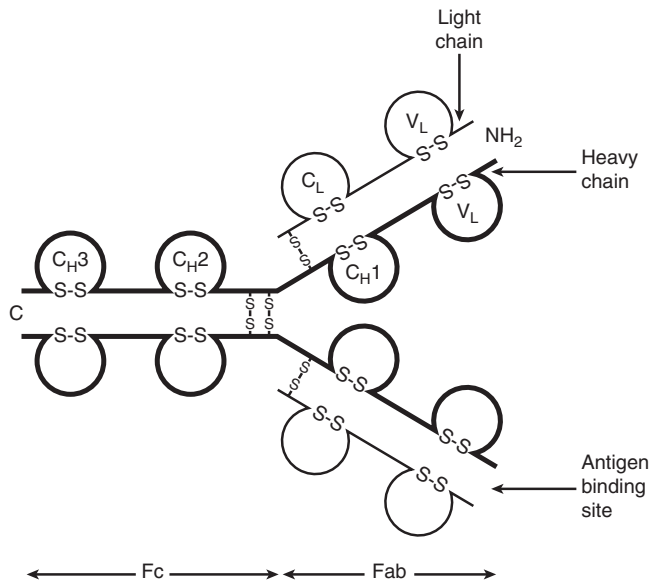


FIGURE 30.5 ■ The basic structure of an immunoglobulin molecule.

The class (or subclass) of the heavy chains and type of light chains is determined by the common amino acid sequences in areas of the molecule called the constant domains. The light chains have one constant domain and the heavy chains have three or four constant domains, depending upon the class; it is these areas that confer functionality to secreted immunoglobulin molecules. Each chain also has a single domain of great variability in the amino acid structure, called the variable domain. The heavy and light chains are arranged such that their

variable domains are adjacent, forming a groove in which antigen binding occurs.

If every antigen recognition receptor on B and T cells were encoded on a one gene/one enzyme principle, these receptors alone would require more than the total DNA present within individual cells. The enormous scope for receptors to recognize antigens is achieved by a mechanism of genetic recombination.

In humans, the gene for the production of the immunoglobulin heavy chain is located on chromosome 14. In the germ line, there are large numbers (estimated 250–1000) of variable (V) gene segments followed by small numbers of D (~ 12) and J (~ 4) gene segments. Finally, there is the constant (C) gene segment that can encode for any of the immunoglobulin heavy chains. A diagram of the immunoglobulin heavy chain gene is shown in Figure 30.6. Processing of this genetic material occurs by deletion of the introns (the non-coding DNA) and splicing together of the exons (the coding portions of the gene). One of the many V-region genes is selected and joined to one of the D-region genes, which is joined to one of the J-region genes. This VDJ sequence is then joined to the constant region genes, typically starting with the μ gene, which is the first downstream heavy chain gene in the sequence, resulting in the production of IgM. The genes for the κ and λ light chains are located on chromosomes 2 and 22, respectively, and are similar to the heavy chain gene except that they do not have a D region. Diversity of the variable region is generated by the large number of possible combinations between the V, D and J or V and J regions of the genes, variation in how these genes are spliced and by somatic mutation generated during cell development.

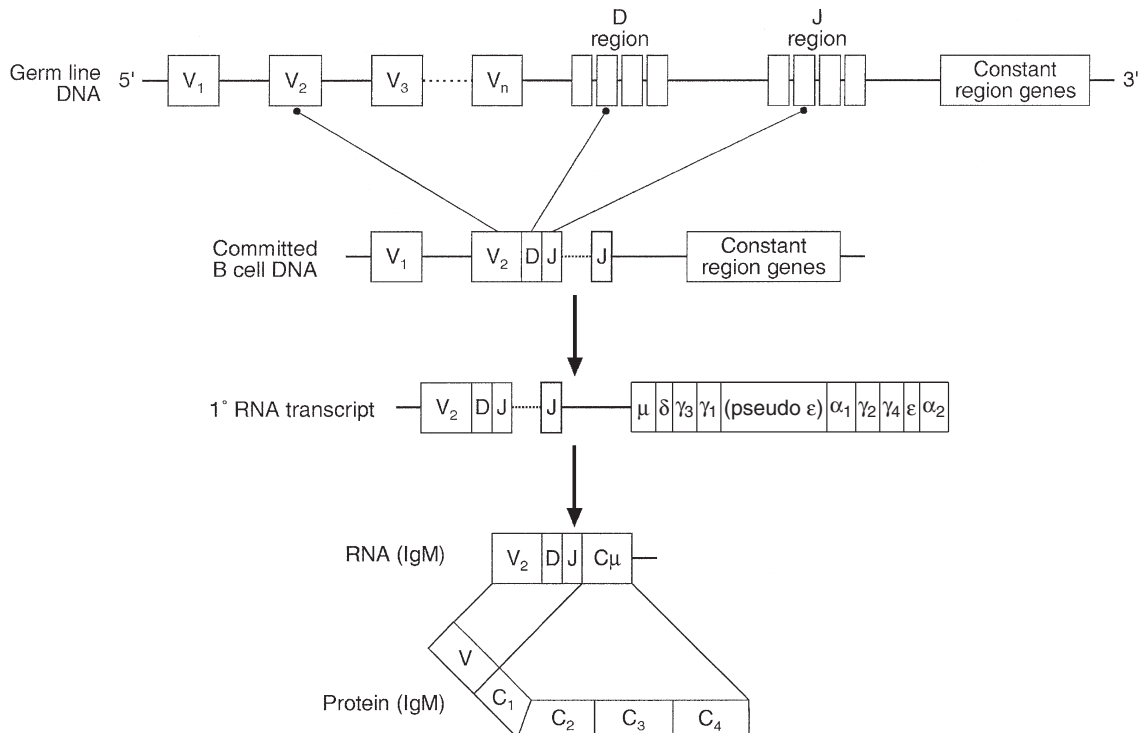


FIGURE 30.6 ■ The immunoglobulin heavy chain gene.

In any cell, there is allelic exclusion, so that only one of each pair of chromosomes coding for immunoglobulin is translated. Each cell, theoretically, has two chances for each gene to be successfully rearranged. The products of unsuccessful rearrangements are secreted by the cell and degraded or cleared by the kidneys. The immunoglobulin heavy chain gene(s) is rearranged first, followed by κ -gene rearrangement. In about 40% of cells, the κ rearrangements are abortive and the cell then proceeds to λ -gene rearrangement. A small excess of free light chains (polyclonal) is produced. The immunoglobulin molecule is then assembled and expressed on the surface of B cells. The cell is now committed to a light chain type and unique variable region specificity (also called the idiotype). The organization of the heavy chain gene enables the variable region to be spliced onto a different heavy chain by the process of 'class switching'. This occurs after interaction with antigen, when cells that show appropriate recognition of the antigen are driven to proliferate. IgM is the antibody produced in primary immune responses; IgG (or IgA or IgE), produced after class switching, is produced in secondary immune responses, as shown in Figure 30.7. Some cells will mature into memory cells and others into end-differentiated plasma cells. Plasma cells are typically non-proliferative and have little membrane-bound immunoglobulin but secrete large amounts.

IgA and IgM both occur as oligomers of the basic unit. Combination is facilitated by the addition of a short

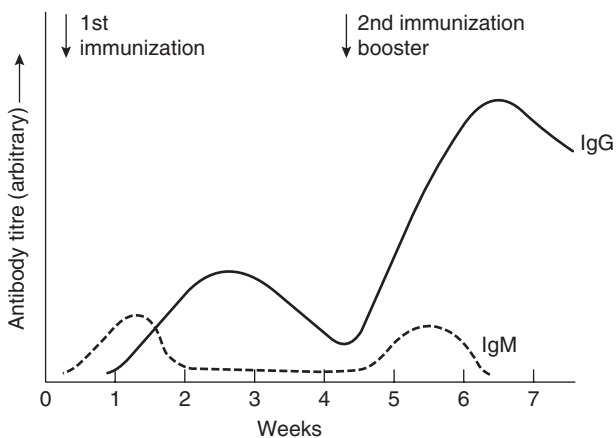


FIGURE 30.7 ■ Antibody production in primary and secondary responses.

polypeptide chain called the J-chain. This J-chain should not be confused with the J gene segments coding for part of the variable region of the immunoglobulin molecule – the two bear no relationship. In the plasma, IgM occurs as a pentamer. IgA can occur as a dimer and this is the predominant form in secretions. A glycoprotein called 'secretory piece' is synthesized by mucosal epithelial cells and integrated into secreted IgA molecules, making them more resistant to degradation by the mucosal environment.

The major properties of the immunoglobulins are shown in Table 30.5.

Under normal circumstances, there is enormous microheterogeneity within immunoglobulin molecules. This results in polyclonal immunoglobulins, representing all classes, light chain types and a large number of different idiotypes, each produced by a different B cell clone, even to a single antigen. Monoclonal immunoglobulins, which consist of a single heavy chain class, light chain type and idiotype, result from the proliferation of a single clone of B cells. In fact, polyclonal immunoglobulins comprise very small amounts of many thousands of monoclonal immunoglobulins.

Each class and subclass of immunoglobulin has its own function or group of functions. IgG is the most abundant antibody in the circulation and also diffuses into the extravascular compartment bathing all the tissues. It is produced in secondary immune responses and is vital for immunological memory. There are four subclasses, with IgG1, IgG2, IgG3 and IgG4 being approximately 65%, 25%, 7% and 3%, respectively, of the total IgG. IgG1 and IgG3 responses mature earliest, while IgG2 and IgG4 responses mature more slowly. IgG1, IgG2 and IgG3 are all potent activators of the classical complement cascade. IgG (bound to an antigen) binds to the $Fc\gamma$ receptors I, II and III on neutrophils, macrophages etc. to facilitate and enhance phagocytosis. IgG has a half-life of approximately 21 days, but this depends upon the plasma concentration, catabolism being increased with raised IgG concentrations and decreased with low IgG concentrations.

IgA is the major immunoglobulin in mucosal secretions and is also produced in secondary immune responses. It is most important in binding and neutralizing organisms at mucosal surfaces, without interacting with other immunological systems.

IgM is the antibody of the primary immune response, culminating in the process of class switching, inducing

TABLE 30.5 The major properties of immunoglobulins

Property or function	IgG	IgA	IgM	IgD	IgE
Molecular weight (kDa)	150	160 (monomer)	900 (pentamer)	185	200
Adult plasma reference range (g/L)	6.0–16.0	0.8–4.0	0.5–2.0	0.01–0.5	<0.1
% Carbohydrate	3	8	12	13	12
Subclasses	4	2	0	0	0
Complement fixation: classical pathway	Yes	No	Yes	No	No
Complement fixation: alternative pathway	No	Yes	No	No	No
Crosses placenta	Yes	No	No	No	No
Binds to Fc receptors on phagocytes	Yes	No	Yes	No	No
Binds to Fc receptors on mast cells	No	No	No	No	Yes

the B cells to make the antibodies of the secondary responses, IgG or IgA. The large pentameric structure of IgM gives good potential for antigen binding and its high molecular weight keeps it restricted to the vascular compartment. The half-lives of IgA and IgM are independent of their plasma concentrations and are approximately five days for both.

IgD is detectable on B cells early in their development, but its concentration in plasma is normally very low. Its exact function is not known, but it is likely that it is important in processing antigens in immature B cells.

The plasma concentration of IgE is very low; it relies on its binding to the F_ε receptor on mast cells and basophils for its activity. The release of inflammatory mediators within these cells is precipitated by binding of the surface-bound IgE to a specific antigen or allergen.

T cell receptors. The T cell receptor is somewhat simpler. It is a heterodimer in which each polypeptide chain is composed of two domains, one variable and one constant. There are four possible polypeptide chains, α, β, γ and δ. A receptor is made up of a pairing of either α with β or γ with δ. In no cell are both types expressed, nor are other pairings possible (e.g. α with δ). The variable domains are similarly constructed from gene segments, as described for immunoglobulin receptors (V, D and J for β and δ; V and J for α and γ), and then spliced into the corresponding constant regions.

The T cell receptor differs from the immunoglobulin receptor in that it has no hypermutable sequences and therefore does not show affinity maturation. Also, it can only recognize antigen that is bound to an HLA (human leukocyte antigen) molecule expressed on the surface of an appropriate antigen presenting cell (APC).

Human leukocyte antigens (HLA). The major histocompatibility complex (MHC) is a group of molecules that are integral to the development of adaptive immune responses. In humans, the genes for these molecules, which are designated human leukocyte antigen (HLA), are on chromosome 6. They are expressed on the cell surface and bind peptides in an antigen-binding groove, where they are recognized by T cells via the T cell receptor. There are two types of HLA, class I and class II, summarized in Table 30.6.

HLA molecules bind antigen but they are different from the receptors on B and T cells in that there is no genetic re-arrangement. Instead, the HLA genes are highly polymorphic, i.e. their nucleotide sequences vary between individuals giving different alleles, summarized in Figure 30.8.

HLA class I molecules are widely expressed on most nucleated cells. They comprise a small non-membrane polypeptide, β₂-microglobulin (the invariant chain of MHC class I) covalently linked to a heavy chain, an integral membrane protein with external domains. The variation within MHC class I arises from polymorphisms within genes for the heavy chains. There are three loci for class I (HLA-A, HLA-B and HLA-C). Each individual inherits A, B and C from each parent and, unlike the situation with immunoglobulins and the TCR, there is no allelic exclusion: both sets of genes are translated. All types

TABLE 30.6 Comparison of HLA class I and class II

	Class I	Class II
Structure	Variable heavy chain + invariant chain (β ₂ -microglobulin)	Two variable chains, α and β
Expression	Most nucleated cells	Antigen presenting cells; may be induced on other cell types
Antigen presented	Intracellular antigen, i.e. viral, self 8–9 amino acids	Extracellular antigen, e.g. microbes or their products 14–22 amino acids
Gene loci	HLA-A, B, C	HLA-DQ, DR, DP
Responding T cell	CD8+ T cells	CD4+ T cells

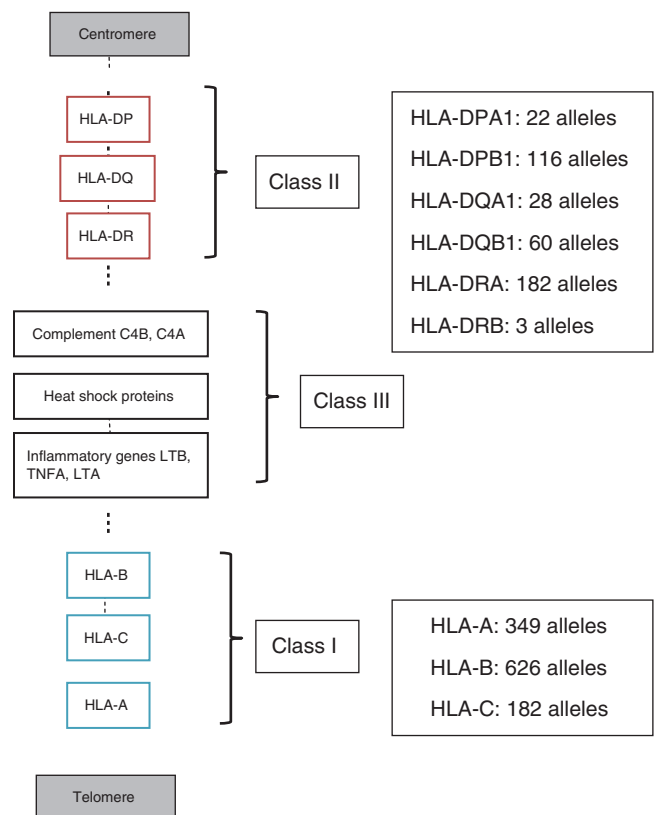


FIGURE 30.8 ■ Simplified overview of the MHC locus on human chromosome 6. The MHC locus is 4MB and encodes ~ 300 genes. Dashed lines are used to indicate that other genes lie between the ones shown.

occur on any one cell. Because of the polymorphisms, it is likely that any individual will be heterozygous at each locus; therefore six products will be formed.

HLA class II molecules are primarily expressed on immune cells, particularly those able to present antigen. They may be inducible on many other cell types. They comprise two transmembrane polypeptides (α and

β chains) of similar size. As for class I, there are three loci (HLA-DR, HLA-DQ and HLA-DP). Again, the likelihood is that these properties will result in at least six distinct receptors. The class II system is, however, more complex. During evolution, there has been replication of some loci such that it is possible to express a number of different DR and DQ types, leading to increased diversity. Even with the variations described above, it is apparent that the total number of different HLA types expressed in any one individual is relatively small compared with immunoglobulins and TCRs.

Antigen presentation

Unlike immunoglobulin and the TCR, HLA molecules are loaded with peptide antigen intracellularly during molecule assembly, so that prior 'loaded' HLA is expressed on the cell surface. Indeed, it would appear that antigen-binding is essential for stability of the molecule and that HLA without antigen is intrinsically unstable. Antigen binds in grooves formed by the folding of class I molecules and the juncture of the α and β chains in class II. The source of the antigen is different for each class. Class I molecules bind antigens synthesized intracellularly, e.g. self-peptides or viral antigens; if these antigens are able to stimulate an immune response it is via CD8⁺ T cells, resulting in destruction of the target cell. Class II molecules bind antigen that was derived extracellularly, e.g. microbes or microbial products. These antigens are internalized within the APC, where they are subjected to partial proteolytic degradation and then bound to the class II molecules for antigen presentation. These antigens are recognized by CD4⁺ T cells and stimulate an immune response.

B cells can act as antigen presenting cells. Their surface immunoglobulins are able to react with free antigen molecules, forming complexes that are taken into the cells, where the antigen is degraded into peptides. These peptides can combine with HLA class II molecules to allow antigen presentation to T cells.

Cellular immune activation

Antigen binding alone will rarely stimulate an immune response. Some very repetitive antigen sequences, for example polysaccharides, may directly stimulate B cells. To respond to most antigens, however, both T and B cells require additional signaling from cytokines or contact with other cell types mediated through accessory or co-ligatory molecules expressed on these cells. Accessory cell interactions are many and complex and the reader is referred to dedicated immunological texts for an account of these.

Complement

The complement system

The complement system consists of approximately 30 plasma proteins, normally present in inactive forms (zymogens). Activation results in the generation of a cascade of enzymes that leads to activation of the terminal cytolytic pathway (membrane attack complex). Activation can occur via two major pathways, the

classical pathway and the alternative pathway. In addition, complement can be activated through the lectin pathway and may be recruited by proteolytic activity generated by other physiological pathways such as coagulation. The system is inherently labile and subject to regulation by various proteins (see below). A simple diagram of the complement system is shown in [Figure 30.9](#). The role of the complement system as a whole is to destroy invading pathogens via the membrane attack complex and opsonization but there are many other effects of complement activation, as shown in [Table 30.7](#). During activation of the earlier components (C2, C4, C3 and C5), smaller peptides are cleaved off the zymogens. These have powerful biological activities: they are able to initiate and potentiate inflammation by increasing blood vessel permeability and attracting inflammatory and immune cells to a site of injury via a process termed chemotaxis.

The larger fragment formed on activation usually has the alphabetical annotation 'b' (e.g. C3b) and the smaller fragment, 'a' (e.g. C3a). Some authors use a line over a component name (e.g. C1 \bar{s}) to denote an active enzyme that can activate the next component; this convention is unnecessary and has not been adopted in this chapter.

Activation via the alternative pathway

The alternative pathway is initiated at the level of C3, bypassing the earlier components C1, C4 and C2, which are activated in the classical pathway. C3 spontaneously degrades in tissue fluids and is rapidly inactivated. However, on activating surfaces, including damaged tissue and components of bacterial cell walls such as lipopolysaccharides, degraded C3 is stabilized and can form an enzyme to activate further C3 and to generate an enzyme to cleave C5 and thus activate the terminal sequence of the pathway.

Activation via the classical pathway

Complexes of antigens with their antibodies (immune complexes) binding C1 activate the classical pathway. C1 is a trimolecular complex of C1s, C1r and C1q. Binding of C1 liberates an enzyme C1s (the C1 esterase), which is able to cleave C4 and C2, and an enzyme formed from a complex of the larger fragments of these two molecules can activate C3. The pathway beyond this is common with the alternative pathway.

Activation via the lectin pathway

Various components of bacterial cell walls are able to bind to a protein in the blood called mannose binding lectin (MBL); this complex is able to stimulate serine proteases that can directly activate component C4.

Regulation of the complement pathways

Many of the complement proteins have a natural cleavage site, making them inherently unstable; therefore the presence of a number of regulatory proteins is vital to stop uncontrolled activation of the cascade. Some of the control proteins circulate while others are membrane

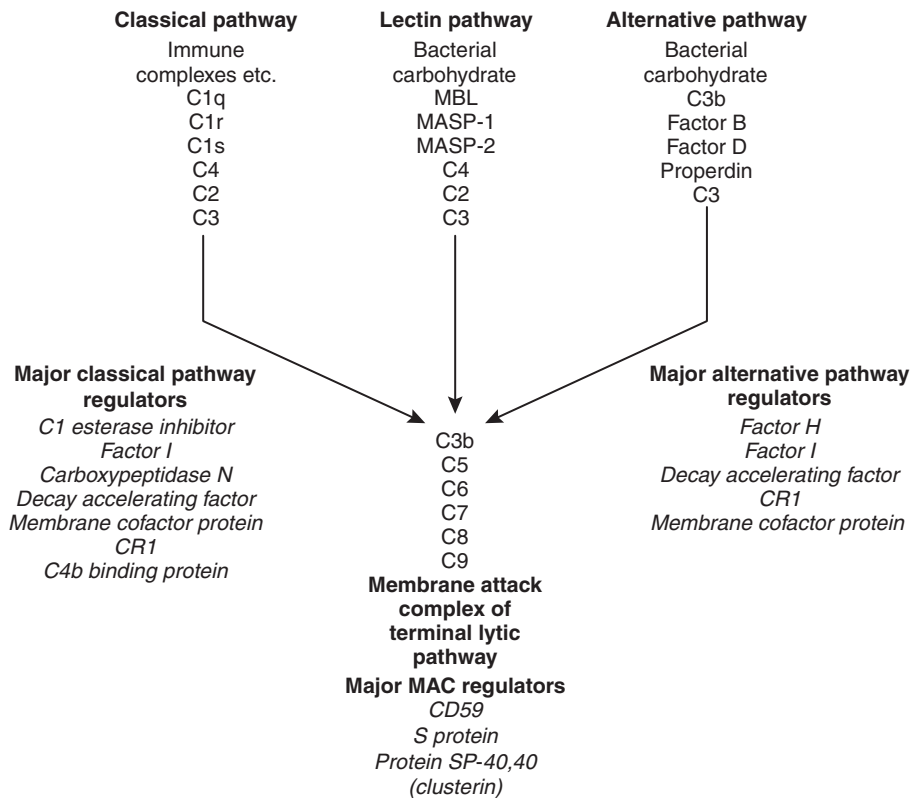


FIGURE 30.9 ■ The complement system, showing the pathways for activation and their major regulatory proteins.

TABLE 30.7 The major actions of components of the complement system

Complement components	Effect
C1q, C3, C4	Clearance of immune complexes Clearance of cell debris
C3a, C4a, C5a (anaphylatoxins)	Histamine release from mast cells Increase smooth muscle contraction Increase vascular permeability Inflammation
C3b (opsonin)	Binding of antigen–antibody complex to receptors on phagocytes, enabling effective ingestion and destruction
C5a, C3a (chemotaxins)	Recruitment of phagocytes into areas of inflammation
C5–C9 (membrane attack complex)	Cell or bacterial lysis Cell signalling

bound; some have enzyme activity, others are simple binding proteins. The important complement regulatory proteins are shown in [Table 30.8](#).

Acute phase proteins

Inflammation is a fundamental pathological process that occurs in response to tissue injury. The inflammatory response has three main stages: local reactions; destruction and/or removal of the injurious material, and repair and healing of the damaged tissue. Stimulators of

TABLE 30.8 The major control mechanisms of the complement cascade

Complement control mechanism	Activity
C1 esterase inhibitor	Protease inhibitor that blocks the activity of C1 esterase
Factor H	Binds C3b and enhances the degradative action of Factor I
Factor I	Enzyme that degrades C3b and C4b
C4b binding protein	Binds to C4 and enhances its destruction by Factor I
Protein S and SP-40,40 (clusterin)	Binds C5b67 and prevents formation of the membrane attack complex
Carboxypeptidase N	Enzyme that inactivates C3a, C4a and C5a
Decay accelerating factor	Transmembrane glycoprotein on most blood cells that binds C4b and inhibits C3 convertase
Membrane cofactor protein (CD46)	Membrane-bound C3 binding protein
Membrane attack complex inhibitory factor (CD59)	Membrane-bound protein that hinders insertion of the membrane attack complex
Complement receptor I (CRI, CD35)	High affinity receptor for C3b and C4b on surface of erythrocytes

inflammation include trauma, infection, infarction and deposition of immune complexes; these result in the release of inflammatory mediators (see [Table 30.9](#)) and the classic signs of inflammation – redness, swelling, warmth or heat, pain and loss (or impairment) of function. These

TABLE 30.9 Inflammatory mediators and their actions

Mediator	Action
Platelet activating factor, histamine, prostaglandins	Vasodilatation
Platelet activating factor, histamine, complement C3a and C5a, bradykinin, leukotrienes	Increased vascular permeability
IL-8, IL-1, and TNF α , complement C5a, leukotrienes	Leukocyte adhesion and chemotaxis
Bradykinin and prostaglandins	Pain
Enzymes (proteases) and products of neutrophil respiratory burst (free radicals)	Tissue damage
Inflammatory cytokines IL-1, IL-6 and TNF α	Hepatic production of acute phase proteins

features may be localized, but in more severe inflammation there is a widespread response and the features also include fever, leukocytosis and the production of a variety of liver-derived acute phase proteins. The functions of these proteins are summarized in [Table 30.10](#).

The acute phase proteins show diverse properties, particularly in the kinetics of their responses, as shown in [Table 30.11](#). The magnitude of response is generally related to the degree of injury, although other factors such as catabolic rate, hormonal influences and genetic variability are also important.

Cytokines

These are biological signalling peptides, whose many actions include the control of the differentiation of white blood cells and the modulation of the actions of the cells of the immune system. They include the interleukins, interferons, tumour necrosis factors and various growth factors. They are summarized in [Table 30.12](#).

In general, cytokines, like hormones, act through high-affinity cell surface receptors that can be widely distributed on many cell types or can be restricted to one or two cell types. The receptors can be up- or downregulated or induced, depending upon the prevailing conditions. Most cytokines are glycosylated and their binding to the receptors may be influenced by the configuration of the sugar residues and by the other cytokines being released in the vicinity. Most cytokines are multifunctional

TABLE 30.10 The functions of acute phase proteins

Acute phase function	Examples
Mediating	Working as parts of networks of inflammatory mediators: C-reactive protein (CRP) binds a variety of ligands and activates complement; complement components are important opsonins and chemotactic factors; fibrinogen and the clotting components form clots and fibrin matrices as a basis for repair
Inhibiting	Inhibiting protease activity and controlling pathways: α_1 -antitrypsin and α_1 -antichymotrypsin inhibit the actions of enzymes released from leukocytes during phagocytosis; C1 esterase inhibitor inhibits part of the complement system
Scavenging	Inhibiting or eliminating noxious substances produced during the inflammatory process: haptoglobin combines with free haemoglobin to form a complex that is rapidly cleared by the liver; CRP may opsonize DNA and cell membrane debris
Regulating	Modulating the immune response: α_1 -acid glycoprotein is expressed on lymphocyte cell membranes
Repairing	Control and laying down of connective tissue elements: α_1 -antitrypsin and α_1 -antichymotrypsin are deposited in a sequential fashion on the surface of newly formed elastic fibres; α_1 -acid glycoprotein promotes fibroblast growth

molecules with diverse biological actions upon a wide variety of target cells. Cytokines, in general, show significant overlap in their functions, with any one function typically being shown by a number of cytokines (pleiotropism). They mostly act locally in either an autocrine or paracrine fashion; very few, for example IL-6, show a true endocrine function. The major differences between cytokines and hormones are shown in [Table 30.13](#). Genetic deficiencies of cytokines are rare, possibly reflecting their vital role in the maintenance of health.

TABLE 30.11 Properties of the major acute phase proteins

Protein	Molecular mass (kDa)	Half-life (days)	Normal adult plasma concentration (g/L)	Magnitude of increase in concentration
Caeruloplasmin	135	2.5	0.2–0.6	} Up to 50% increase
Complement C3	185	2.5	0.7–1.7	
α_1 -Acid glycoprotein	41	2.3	0.5–1.0	
α_1 -Antitrypsin	54	4	1.4–3.2	} 2–3-fold increase
α_1 -Antichymotrypsin	68	2	0.3–0.6	
Haptoglobin	86	4	0.5–3.2	
Fibrinogen	340	4–6	2.0–3.5	} Up to 1000-fold increase
C-reactive protein	105	1	<0.005	
Serum amyloid A	12	1	<0.001	

TABLE 30.12 The major groups of cytokines

Cytokine	Major cell sources	Major immunological functions
Interleukins (ILs; over 35 interleukins have now been characterized)		
IL-1 α and β	Monocytes Macrophages Endothelial cells Fibroblasts	Activation of T and B cells, macrophages and endothelium Acute phase response and production of acute phase proteins Fever
IL-2	T cells	Activation and proliferation of T, B and NK cells
IL-4	T and B cells Macrophages mast cells	Activation of B cells Differentiation of Th2 cells Suppression of Th1 cells
IL-5	T cells Mast cells	Eosinophil recruitment, activation and development
IL-6	Monocytes Macrophages T cells Endothelial cells Fibroblasts	Activation of stem cells Differentiation of T and B cells Acute phase response and production of acute phase proteins
IL-8	T cells Monocytes Neutrophils Endothelial cells Fibroblasts	Chemoattraction and activation of neutrophils Chemoattraction of T cells and basophils
IL-10	T and B cells Macrophages	Suppression of macrophages Suppression of Th1 cells Activation of B cells
IL-12	B cells Macrophages Dendritic cells	Differentiation of Th1 cells Activation of T and NK cells
IL-13	T cells	Activation of B cells Suppression of Th1 cells
IL-15	Macrophages	T and B cell proliferation
IL-17	Activated T cells	Inflammation
IL-18	Macrophages	IFN γ production by T cells
IL-23	Macrophages Dendritic cells	IL-17 production
IL-35	Regulatory T cells	Suppresses T cell proliferation
Interferons α, β, γ		
IFN α, β	T and B cells Monocytes Macrophages Fibroblasts	Antiviral Stimulation of macrophages and NK cells Upregulation of HLA class I expression
IFN γ	T cells NK cells	Antiviral Stimulation of macrophages and endothelium Upregulation of HLA class I and class II expression Suppression of Th2 cells
Tumour necrosis factor superfamily (>50 members)		
TNF	Macrophages T and B cells Neutrophils Endothelium Mast cells	Anti-tumour Activation of macrophages, endothelium and granulocytes Upregulation of HLA class I Acute phase response and production of acute phase proteins Weight loss (cachexia)
BAFF ^b	Dendritic cells T cells	Enhancement of expression of B cell survival factors
Colony stimulating factors (G-, M-, GM-CSF)		
G-CSF	T cells	Activation and development of neutrophils
M-CSF	Macrophages Neutrophils Fibroblasts Endothelium	Activation and development of monocytes and macrophages
GM-CSF	T cells Macrophages Neutrophils Mast cells Eosinophils	Stem cell differentiation Development of neutrophils, macrophages and eosinophils
Growth factors (GF) Chemokines	Fibroblast, platelet derived, epidermal, insulin like, transforming, erythropoietin e.g. lymphotactin, monocyte chemotactic protein-1, IL8, RANTES ^a , macrophage inflammatory protein 1 α	

^aRANTES: regulated on activation, normal T expressed and secreted; Th, T helper.^bBAFF: B Cell activating factor.

TABLE 30.13 Major differences between cytokines and endocrine hormones

Property	Cytokine	Hormone
Sites of production	Many and varied	Many, but specific to each hormone
Cellular targets	Few	Many
Biological role	Fighting infection, tissue repair	Homoeostasis
Biological redundancy	High	Low
Biological pleiotropy	High	Low
Present in the circulation	Rarely	Yes
Sphere of influence	Predominantly autocrine/paracrine	Widespread and distant from site of production
Inducers	External insults	Primarily physiological changes (except stress hormones)

Many cytokines interact with each other and the effect of a cytokine may change depending upon the prevailing micro-environment. It is important to stress that cytokine synthesis, including that of the proinflammatory cytokines, is a normal response to injury. Their functions should not be thought of as exclusively harmful, resulting in tissue damage or systemic shock. In the majority of instances, cytokines act to coordinate the elimination of invading organisms or removal of damaged tissue, thus avoiding excessive stimulation of specific immunity that could lead to hypersensitivity reactions.

Inflammatory cytokines

The cytokines interleukin (IL)-1 β , IL-6 and tumour necrosis factor alpha (TNF α) are all central to the initiation of the inflammatory response. However, they do not function in isolation: other cytokines implicated in the inflammatory response include IL-8, 10, 11, 12, 17, 18, 23 and interferon γ . In addition, both defensive and injurious effects will involve the release of many mediators of inflammation such as peptides, for example the complement fragment C5a, and lipids, for example platelet activating factor. Many of these mediators act synergistically and also induce each other's production and the production of other cytokines, resulting in both positive and negative feedback control pathways. Cytokines are also responsible, either directly or indirectly, for healing and successful resolution of the inflammatory insult. Severe inflammatory responses frequently overlap with specific immune responses, as they are often induced by infection and, furthermore, the tissue damage accompanying severe trauma may lead to secondary infection.

Mechanisms of immunological damage

Immune responses are usually well controlled and host damage is minimal and reversible. Under some circumstances, resolution does not occur and an exaggerated or

persistent response leads to irreversible host tissue damage and, in extreme situations, even death. These inappropriate reactions are termed hypersensitivity reactions. These are classified into four types (I-IV), each with a different mechanism. Type I hypersensitivity reactions, central to allergic responses, may occur without involvement of any other type of hypersensitivity reaction. Most reactions, however, involve more than one type of response and do not, alone, define a pathogenetic mechanism that underlies a group of diseases. There are also other reactions that do not fit neatly into this classification, for example stimulatory reactions, such as when specific immune cells are stimulated by external agents, and when autoantibodies have a stimulatory effect, for example as occurs with thyroid receptor antibodies in Graves disease.

Hypersensitivity reactions result in the release of inflammatory mediators. Some mediators have direct effects on local, or even distant, tissues; others recruit and activate effector cells that further contribute to tissue damage.

Type I hypersensitivity

Type I hypersensitivity reactions are IgE mediated. The IgE antibodies are formed to an antigen (or allergen), with an individual's tendency towards making IgE being determined by many factors including genetic, T cell responsiveness and antigenic burden. The IgE binds to high-affinity IgE receptors on the surfaces of mast cells and basophils, and these cells are now primed to react the next time the cells come into close proximity with the allergen. The cross-linking of IgE on the cell surfaces causes rapid cellular degranulation and liberation of a number of chemical mediators. The mediators released by mast cell degranulation include the preformed molecules histamine, protease enzymes, proteoglycans (heparin) and chemotactic factors. Reaction of antigen with IgE on mast cells also stimulates synthesis and release of platelet activating factor (PAF), leukotrienes (B₄, C₄ and D₄) and prostaglandins (mainly PGD₂). The mediators of type I hypersensitivity reactions are shown in Table 30.14.

The actions of histamine depend on the site of release. In the airways, it induces smooth muscle contraction; in the skin, it causes the hallmark wheal and flare response. Widespread activation of mast cells leads to systemic effects

TABLE 30.14 Mediators of type I hypersensitivity reactions

Mediator	Pharmacological effect
Histamine	Vasodilatation, capillary permeability, bronchoconstriction
Heparin	Control of histamine release
Leukotrienes (various)	Bronchoconstriction, airway tissue oedema
Prostaglandins (various)	Potent mediators of inflammatory response
Platelet activating factor	Platelet aggregation
Tryptase	Proteolytic enzyme activates C3
Kininogenase	Kinins \rightarrow vasodilatation \rightarrow oedema
Cytokines (IL-5, IL-8, TNFs)	Chemoattractants

of circulatory shock, hypotension, collapse, chest tightness and, in the most severe cases, respiratory arrest and death: this is anaphylactic shock. Type I hypersensitivity reactions occur rapidly (within approximately 20 min of an insult) and are also called 'immediate hypersensitivity reactions'.

Type II hypersensitivity

The important characteristic of type II hypersensitivity reactions is that the antigens involved are localized to the membranes of the target cells; they may be synthesized cell products, cell membrane components or bound foreign molecules (e.g. a drug). The circulating antibodies (IgG or IgM) bind to the cell-bound antigen and activate complement. Full activation of the complement pathway results in the membrane attack complex (C5–9) being assembled on the cell surface, resulting in target cell lysis. Partial activation via C3 will also render the cell a target for phagocytosis, via C3b and IgG that bind to appropriate receptors on phagocytic cells. Activation of phagocytic cells leads to release of further tissue-damaging enzymes. The tissue damage persists for as long as antibody and antigen are present.

Some of the more important type II responses involve red blood cell antigens (including transfusion reactions and haemolytic disease of the newborn) and autoantibodies to cell surface components, for example autoimmune haemolytic anaemia (in which the target is red blood cells), idiopathic thrombocytopenia purpura (in which the target is platelets) and Goodpasture syndrome (in which the target is the kidney glomerular basement membrane).

Type III hypersensitivity

Type III hypersensitivity reactions are also termed immune complex reactions. Complexes of antigen and antibody form in the circulation and are then deposited in susceptible tissues; they may also form directly in the tissue. The latter mechanism is termed the Arthus reaction, and is typically seen with repeated insect stings, where a red swollen lesion develops after a sting. The tissue damaging mechanisms are similar to those described for the antigen-antibody complexes that form in type II responses. The response times of types II and III hypersensitivity reactions are slower than that of type I reactions; they typically develop 3–6 h after exposure to antigen. The response can also become chronic, particularly in autoimmune reactions, where antigen persists.

The clinical manifestations of type III hypersensitivity reactions relate to the tissue deposition, for example vasculitic (skin), serum sickness (systemic), nephritis (kidneys) and extrinsic allergic alveolitis (lungs).

Type IV hypersensitivity

These reactions are cell mediated. Membrane receptors on sensitized T lymphocytes recognize antigen and the cells (particularly Th1) release cytokines that in turn activate macrophages. The major mediators of the hypersensitivity response are products of activated macrophages including enzymes, coagulation factors, complement, superoxide ions, leukotrienes, prostaglandins and cytokines. The hallmark of responses to

persistent antigen is formation of granulomas that wall off the inflammatory focus.

Type IV reactions are also known as delayed type hypersensitivity reactions because symptoms occur 24–48 h after re-exposure to antigen. They are seen in response to mycobacteria (e.g. *Mycobacterium tuberculosis*) and are the mechanism underlying contact dermatitis.

Conclusion

This brief introduction has described the major components of the immune system and their interactions during immune responses. Antigens, typically encountered across mucosal surfaces, are taken up by antigen presenting cells (APCs) and processed and presented on the major histocompatibility antigen (MHC) molecules of these cells to the lymphocytes in lymphoid tissue. The lymphocytes, with their randomly rearranged receptors (TCR and Ig), traffic around the lymphoid tissue until activated by an antigen, whose three-dimensional shape can engage the receptor. Activated cells communicate via cytokines and various peptides and proteins to induce lymphocyte proliferation and maturation. Plasma cells of the B lymphocytic lineage secrete antibody that circulates, coats the mucous membranes and bathes the tissues. Re-exposure to a previously encountered antigen stimulates a rapid and specific response via antibody, complement, phagocytosis and T cell killing to provide long-lived immunity to infection. The complexity of the system makes it robust but also provides opportunities for something to go wrong and result in diseases of the immune system.

DISEASES OF THE IMMUNE SYSTEM

Introduction

Diseases of the immune system can be classified as follows:

- immune deficiency resulting from failure of the development or function of immune organs, cells or other components, either from inherited causes (primary immune deficiency) or from some other disease (secondary immune deficiency)
- allergic reactions, which may result from any of the hypersensitivity reactions previously described or from direct activation of inflammatory mediators
- autoimmune disease resulting from immune reactions against body tissues
- malignancy of organs or cells of the immune system.

In addition, severe sepsis resulting from damage to the immune system rather than direct bacterial tissue damage may be considered as an immunological disease.

These disorders will be considered in more detail in the following sections.

Immune deficiency

Development of immunity in humans

The ability to recognize antigen is present at birth, but except on the rare occasions when there has been an intrauterine infection, antigens will not previously have been encountered. The newborn will not have memory

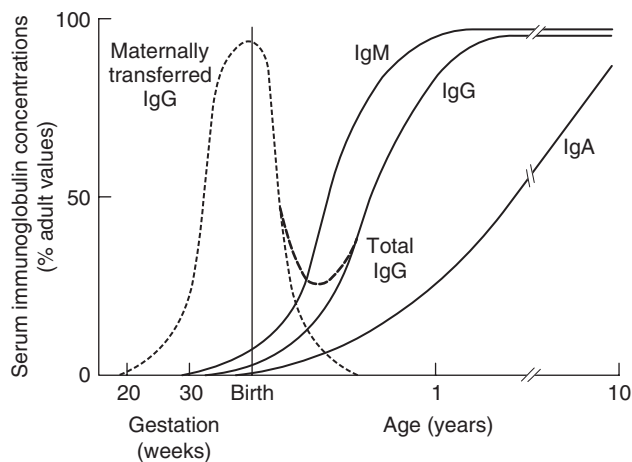


FIGURE 30.10 ■ Serum immunoglobulin concentrations with age. IgG is actively transferred across the placenta to be gradually replaced by the IgG synthesized by the infant. The assays should be capable of reliably detecting concentrations for IgG of 1 g/L, IgA of 0.07 g/L and IgM of 0.1 g/L (after Adinolfi (ed). Immunology and development. Heinemann: London, 1969).

responses so are potentially vulnerable to infection. In the first 3–6 months of life, during which many microbes will be met, immunity is provided by maternally derived IgG, giving time for a child's own system to mature and to establish memory. Active transfer of IgG occurs from 12 weeks of gestation but the major portion is transferred after 32 weeks of gestation. A term baby should have adult plasma IgG concentrations. Premature babies will have reduced IgG concentrations and are more vulnerable to infection. Babies who grow poorly in utero may also have low concentrations. A diagram showing serum immunoglobulin concentrations from pre-term to adulthood is shown in Figure 30.10. Term babies have different immune responsiveness from adults, as shown in Table 30.15.

Deficiencies of the immune system can be due to the lack of a component (or group of components) or defective (or absent) function of a component. Owing to

TABLE 30.15 Immune responsiveness in term neonates

Components	Responsiveness
B cells	Normal numbers but immature (CD5+)
Antibodies	Able to make IgM, with good response to protein antigens but poor response to carbohydrates
Complement	Maternal IgG Classical pathway, 90% of adult; alternative pathway, 60% of adult; C8 and C9, only 20% of adult
T cells	Higher numbers than in adults and immature neonates
Cytokines	IL-2, as adult; IFN γ , 20% of adult; Th2 cytokines very low
Cytotoxicity	Tc only 30–60% of adult; NK 50% of adult

immune interactions, defects in one component may affect another. For example, T cell defects may also result in marked antibody deficiency when T helper activity is compromised.

Infection and immune deficiency

The hallmark of immune deficiency syndromes is frequent, severe or unusual infection. The pattern of infections can give an indication of the type of deficiency, although in young children, this may be difficult to judge. It is estimated that children have eight minor infections per year and it is suggested that significant infection should be regarded as one requiring antibiotic treatment. Even this can be misleading, as the threshold for prescribing antibiotics varies between practitioners. The infections typically seen in specific and non-specific immune deficiencies are summarized in Table 30.16. In general, systemic viral and fungal infections are more typical of T cell deficiencies and common respiratory tract and gut infections are more typical of B cell deficiencies. With regard to the non-specific immune system, defects in complement tend to predispose to most common bacterial infections and impose an increased risk of autoimmune disease, while defects in phagocytes are linked with most common skin infections, cutaneous abscesses and lung abscesses.

In addition to the pattern of infections, there are other factors that can be used to build up an impression of whether immune deficiency should be considered. These are listed in Box 30.1.

Investigation of patients with suspected immune deficiency

The various categorizations of immune deficiency (primary or secondary, cell mediated or humoral etc.) are made at the end of the diagnostic process, not the beginning. The investigation of patients with recurrent, atypical or severe infections, irrespective of age, should follow the same basic pattern, but the detailed investigative process needs to be tailored to the individual patient. A careful history, including the type of symptoms, age of onset of symptoms, organisms responsible (where known) and family history (particularly of immune problems) is essential. Microbiological investigations and basic tests like a full blood count and white cell differential are vital. The process aims to determine the following:

- are all the components of the immune system present?
- are all the components functional (alone or in consort)?
- can the system respond to a defined challenge?
- what is the molecular defect?

As a general guide, the simple tests, and those that are important in the more common immune deficiencies, should be done first. The more complex, functional and genetic tests should be done later in the investigation process (unless there are strong indications from the clinical manifestations or family history). Figure 30.11 shows a basic plan for the immunological investigation of a patient with suspected immune deficiency. Following

TABLE 30.16 Infections typically seen in deficiencies of the immune system

<i>Specific immune system</i>		<i>Non-specific immune system</i>	
<i>T cell deficiencies</i>	<i>B cell deficiencies</i>	<i>Complement</i>	<i>Phagocytes</i>
Common infections			
Viruses: Cytomegalovirus (CMV) Herpes Adenovirus	Bacteria: <i>Staphylococci</i> <i>Streptococci</i> <i>Haemophilus</i> <i>Neisseria</i>	Bacteria: <i>Haemophilus</i> <i>Streptococcus</i> <i>Neisseria</i>	Bacteria: <i>Staphylococci</i> Fungi: <i>Candida</i> <i>Aspergillus</i>
Fungi: <i>Candida</i> <i>Aspergillus</i> <i>Pneumocystis carinii</i>			
Protozoa: <i>Cryptosporidium</i>			
Less common infections			
Bacteria: <i>Mycobacteria</i> <i>Listeria</i> <i>Campylobacter</i>	Bacteria: <i>Campylobacter</i> <i>Salmonella</i> Enteroviruses: Polio Echoviruses	Viruses: Cytomegalovirus Herpes	Bacteria: <i>Salmonella</i>

BOX 30.1**Factors that should be considered in the evaluation of patients with possible immune deficiency**

- Increased number of infections
 - More than three infections requiring antibiotics in one winter
 - Frequent minor episodes
 - More infections than siblings
 - More infections than siblings at same age
- Severe infections
 - More than one severe bacterial infection
 - Frequent requirement for antibiotics
 - Infections that recrudescence when antibiotics are finished
- Unexpected infections
 - Opportunistic infections
 - Unexplained causative organism
 - Unusual infections
- Positive family history
 - Siblings
 - Male relatives on maternal side (for X-linked immune deficiencies)
- Unexplained bronchiectasis
- Therapy-resistant or steroid-resistant asthma
- Severe warts
- Complications of vaccination (e.g. disseminated BCG)
- Failure to thrive
- Chronic diarrhoea

this plan should make it possible to classify the type of immune deficiency.

Primary immunodeficiencies

Clinically significant primary immune deficiency is rare. Many patients suffer years of ill health before they are appropriately investigated and a diagnosis made. The classification of these disorders is by an International System (see [Further reading](#), below) based on the identified defect.

B Lymphocyte (humoral) system

IgA deficiency. This is the most common of the primary immune deficiencies, affecting approximately 1:700 of the population, and is characterized by low plasma IgA concentrations (<0.07 g/L for total IgA deficiency). Patients are often asymptomatic, although they can suffer with repeated respiratory tract and sinus infections, especially if IgA deficiency is associated with IgG subclass deficiency. No particular treatment is indicated, although, if the patient is symptomatic, a low threshold for antibiotic treatment should be adopted. Salivary IgA concentration matures at approximately six weeks of age and can be measured to give an early indication of deficiency. Patients with IgA deficiency can develop antibodies to any IgA present in blood products and this can cause transfusion reactions.

Common variable immunodeficiency (CVID). This is a primary immune deficiency, although the age of presentation can be any time from childhood to late in adulthood. It is not one clearly defined disease, rather a heterogeneous group of undifferentiated syndromes with the common feature of defective (quantitative and/or qualitative) antibody production. The most common presentation is with recurrent sinus and pulmonary infections. The critical laboratory test is quantification of immunoglobulin concentrations (with serum and urine electrophoresis to exclude myeloma). These are typically well below the lower limits of the age-related reference ranges. Patients can also have poor responses to vaccines, low CD4 counts, abnormal cell-mediated immunity and changes in the B cell subset profile. Treatment is with intravenous or subcutaneous immunoglobulin (IgG) replacement, with the dosage being calculated on the basis of body weight. Antibiotics can be given prophylactically or a low threshold adopted for their use in treatment. The IgG should be given every 2–3 weeks and trough immunoglobulin concentrations should be checked regularly, particularly in children.

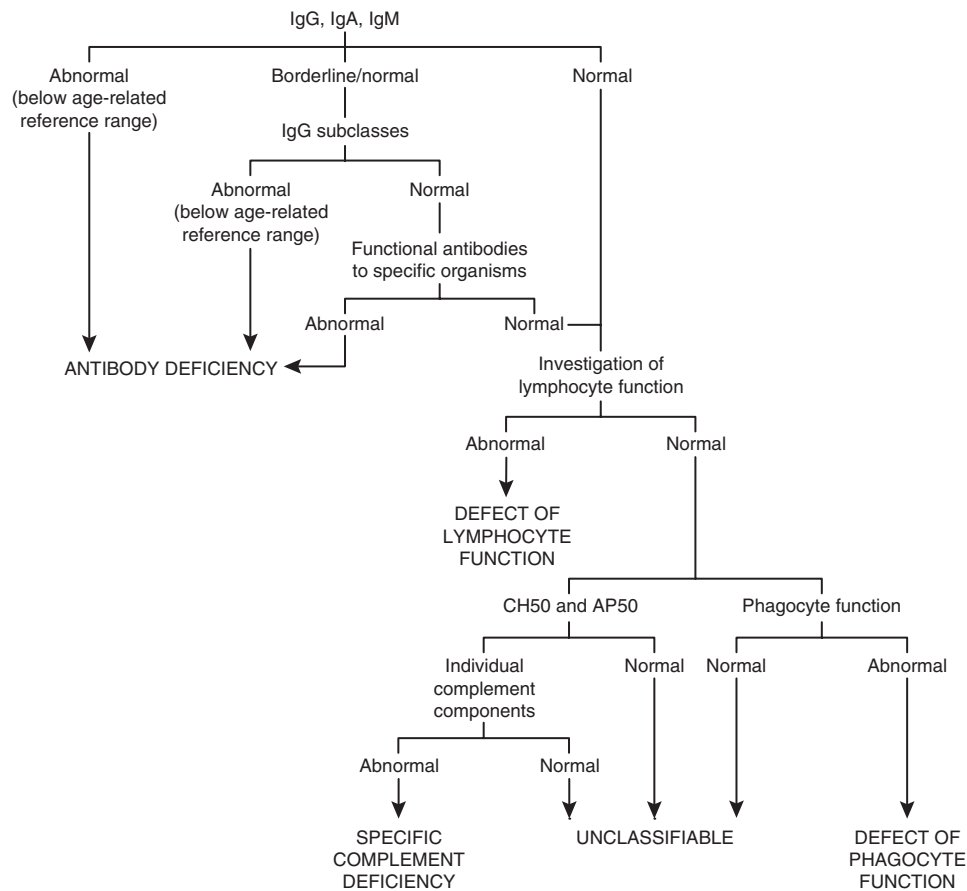


FIGURE 30.11 ■ A basic plan of investigations in suspected immune deficiency (after IUIS/WHO 1981, 1988, 2009).

X-linked agammaglobulinaemia. This condition, also known as Bruton agammaglobulinaemia, is caused by mutations in the gene for a B cell-specific tyrosine kinase (Btk). This results in no B cell development in the bone marrow, a lack of circulating B cells and profoundly low immunoglobulin concentrations. Patients present with recurrent bacterial infection from 4–5 months of age, once protective maternal IgG has disappeared. Treatment is with replacement immunoglobulin, as in patients with CVID.

T Lymphocyte (cell-mediated immunity) system

Severe combined immunodeficiencies (SCID). This is a group of deficiencies with a variety of underlying molecular defects. They are the most profound of the immune deficiency diseases and patients typically present in early infancy. The most common of the diseases is the X-linked SCID that results from a defect in the common γ -chain of the IL-2 receptor. This affects the responsiveness to IL-2, IL-4, IL-7, IL-9, IL-15 and IL-21, so that there is no T cell or NK-cell function; although B cells can be detected, they have no T cell help to make antibodies. The lymphocyte count and T and B cell numbers are the most important investigations when considering SCID; immunoglobulin concentrations are not particularly helpful (owing to the presence of maternal IgG, and the normally low IgA and IgM concentrations expected in young babies). Any baby with suspected SCID should be referred to a paediatric immunologist for investigation. Treatment is with stem cell transplantation or, more recently, gene therapy.

DiGeorge syndrome. This condition is associated with deletions in the chromosomal region 22q11. It is a clinical spectrum of developmental defects including congenital cardiac defects, dysmorphic facies (including cleft palate), immune deficiency secondary to thymic hypoplasia and hypocalcaemia due to parathyroid aplasia or hypoplasia. The abnormal development of the thymus results in reduced T cell numbers and/or function and increased susceptibility to infection. The immunological abnormalities, which are seen in approximately 25% of patients, usually improve with age.

X-linked hyper-IgM syndrome/CD40 ligand deficiency. This immune deficiency is due to a defect in the gene for a cell surface receptor called CD40 ligand (CD154). The binding of CD40 (on antigen-presenting cells and B cells) to CD40 ligand (on T cells) is vital for B cell maturation and class switching. Presentation is similar to X-linked agammaglobulinaemia but these patients are also susceptible to opportunistic infections. They typically have low plasma concentrations of IgG, IgA and IgE and increased concentrations of IgM (as high as 10 g/L). Assessment of the expression of CD40 ligand on T cells is the most important investigation. The bacterial infections can be well controlled with replacement IgG, but patients may develop infections with *Cryptosporidium*, and can develop a sclerosing cholangitis and liver failure. Rare, autosomal recessive forms of hyper-IgM syndrome also exist.

Phagocytic (polymorphonuclear and mononuclear) system

Chronic granulomatous disease. This condition, which can be X-linked or autosomal recessive, is caused by abnormalities in the NADPH oxidase system. This is critical in the respiratory burst and the production of antimicrobial free radicals in neutrophils. Patients present with recurrent abscesses or infections and granulomata. An abnormally low respiratory burst in the neutrophils is the most significant finding; this property is increasingly being investigated by flow cytometric assays. Molecular testing is used to confirm the diagnosis.

Leukocyte adhesion defect types I and II. Recruitment of neutrophils, lymphocytes and monocytes into sites of infection relies on the presence of adhesion molecules, for example CD11a/18 and selectins. Absence or deficient expression of these molecules on cell surfaces results in poor cellular adherence, migration and phagocytosis. Patients present with recurrent, life-threatening bacterial infections and impaired wound healing.

Complement system. Deficiencies of all the complement components have been identified but have varied clinical consequences. Deficiencies of C3 and C6 are particularly associated with recurrent infections. The most important initial investigation is assessment of the integrity of the whole complement system, using haemolytic assays of the classical and alternative pathways or enzyme linked immunosorbent assay (ELISA)-based total complement pathway tests. If these are abnormal, individual components can be investigated: this usually requires referral to specialized centres.

C1 Esterase inhibitor deficiency. This inherited (autosomal dominant) deficiency of the control protein C1 esterase inhibitor results in hereditary angioedema – episodes of subepithelial swelling of the skin (especially of the face, limbs and trunk), gut, larynx etc. The major laboratory findings are low C4 and C1 esterase inhibitor concentrations but normal C3 concentrations. There are rare forms in which antigenic C1 esterase inhibitor concentrations are normal but function is impaired and C4 concentrations are usually low. Measurement of C1 esterase inhibitor should also be considered in patients with apparent anaphylaxis, as the two conditions have similar features in their early stages. However, it should be noted that C4 concentrations may also be low post-anaphylaxis. C1 esterase inhibitor deficiency can also be seen secondary to B cell malignancy when the monoclonal protein interferes with the normal activity of C1 esterase inhibitor.

Transient hypogammaglobulinaemia of infancy. This condition is not strictly a primary immune disorder but is mentioned here for completeness. Some babies show a very slow maturation of their own antibody production, resulting in a lengthening of the physiological trough in concentrations after maternal IgG has disappeared. These babies show normal responses to immunization and will eventually achieve normal antibody concentrations. No particular treatment is indicated, but patients

should be monitored regularly in case antibody concentrations do not improve or they become symptomatic.

Secondary immune deficiency

Secondary immune deficiencies are significantly more common than primary immune deficiencies. They are the predominant form in adults, although they do occur in children. Investigation should concentrate on identifying the underlying process rather than identifying the nature of the defect in immunity. The causes of secondary immune deficiency are summarized in [Table 30.17](#).

There are two broad mechanisms of secondary immune deficiency: inadequate production of a component or excessive loss of a component. Iatrogenic causes of immune deficiency rarely warrant investigation as they are accepted consequences of treatment for other (usually more severe) diseases. Malnutrition is the single most common cause of immune deficiency worldwide.

The most clinically important virally induced secondary immune deficiency is that related to HIV infection. The investigation of this relies particularly on measurements of CD4+ T cell numbers and viral load. Infections with other viruses, for example Epstein–Barr virus and cytomegalovirus, may be associated with some degree of secondary immune deficiency (usually of the cell-mediated responses).

Adults with symptoms of immune deficiency without an obvious cause should be investigated for lymphoid malignancy. These conditions can affect both the cell- and antibody-mediated immunity, making the patient particularly prone to infection. The investigation of lymphoid malignancies is discussed on [p. 589](#).

Protein loss. The most frequent type of immune deficiency associated with excessive component loss is hypogammaglobulinaemia. The loss of immunoglobulins, if sufficiently marked to cause low plasma concentrations, is certain to be accompanied by the loss of albumin

TABLE 30.17 Causes of secondary immune deficiency

Cause	Examples
Immunosuppression	Corticosteroids
Chemotherapy	Cyclophosphamide, methotrexate, vincristine
Radiotherapy	Total body irradiation
Malnutrition	Protein malnutrition, vitamin deficiencies, trace element deficiencies
Infection	HIV, cytomegalovirus (CMV), Epstein–Barr virus, some bacteria, e.g. mycobacteria post-septicaemia
Malignancy	Myeloma, lymphoma, leukaemia
Loss of protein (or other components)	Nephrotic syndrome, protein losing enteropathy, loss from skin, e.g. burns
Miscellaneous	Trauma, type 1 diabetes, splenectomy, inflammatory bowel disease

(a protein of lower molecular weight and higher plasma concentration than the immunoglobulins). It is useful, therefore, to check serum albumin concentrations together with immunoglobulin concentrations when investigating this group of patients. A normal albumin concentration with low immunoglobulins would suggest suppression of production rather than excessive loss. It is also vital to remember that protein loss via the kidneys can be seen with lymphoid malignancy so serum and urine electrophoresis must be performed. Protein loss through the gut can be difficult to prove, but a raised faecal α_1 -antitrypsin concentration is good evidence that protein is being lost into the gut. α_1 -Antitrypsin is used because, as a protease inhibitor, it is relatively resistant to proteolytic degradation in the gut; nevertheless, specimens must be frozen as soon as possible after collection to minimize degradation.

Splenectomy. Patients without a spleen, for example owing to surgical removal following trauma, have an increased susceptibility to overwhelming infection with bacteria such as *Pneumococcus*, *Meningococcus* and *Haemophilus influenzae* type B. Patients who are having elective splenectomy should be vaccinated with pneumococcal, haemophilus and meningococcal A and C vaccines; asplenic patients should take prophylactic antibiotics. The measurement of specific antibody titres (approximately six weeks post-vaccination) may be indicated to confirm adequate antibody responses. Asplenic patients may not maintain optimal antibody concentrations and these should be measured every 2–3 years, depending on the clinical findings. Patients found to have suboptimal concentrations should be re-immunized.

Allergies

Various mechanisms exist, whereby an individual can have an unexpected or inappropriate response to an agent that is not typically considered to be harmful. The mechanisms of such reactions include types I–IV hypersensitivity reactions, direct pharmacological effects (e.g. from biogenic amines in certain foods), biochemical effects (e.g. due to alcohol or other drugs) and intolerances due to enzyme defects (e.g. lactose intolerance). The term *allergy* is often used by lay people to describe such reactions but this term should only be applied to type I hypersensitivity reactions mediated by IgE.

The prevalence of allergies is increasing rapidly; it is estimated that 15–20% of the population have some type of allergy. The clinical features of allergic reactions include:

- abdominal pain
- anaphylaxis
- eczema and atopic dermatitis
- bronchospasm
- conjunctivitis
- diarrhoea
- headache
- malabsorption
- pneumonitis
- pruritus

- rhinitis
- urticaria
- vomiting.

Patients may have only one symptom or a mixture of symptoms. They can range in severity from mild and inconvenient to damaging to health and, in the extreme, to life-threatening anaphylactic shock.

Investigation of patients with allergies

It is important to investigate patients with suspected allergy to provide a logical basis for exclusion of the allergen, with the objective of improving well-being and preventing severe reactions. Investigation must start with the patient's history and examination; the important points in the history are shown in Table 30.18. The exceptions are patients who complain of mild symptoms clearly related to eating one particular food, who often do not need investigation and should simply be advised to avoid that food, and patients with mild reactions to inhaled allergens, for example hay fever and seasonal rhinitis, who should attempt to avoid exposure and be treated symptomatically. However, if a patient has had a severe reaction, investigation is indicated to identify positively the causative agent. Accepted

TABLE 30.18 Important points in taking the history in a patient with suspected allergy

Point	Notes
Age	Babies and young children should be investigated by people with appropriate training. Allergy can have a wider impact in children – loss of schooling, poor health, exclusion from normal activities
Family history	A positive family history increases the likelihood of allergy in an individual
Symptoms (what symptoms and where, e.g. gut, respiratory, skin etc.)	Can help to narrow down the allergen and direct investigation
Symptoms (severity)	Severe symptoms, e.g. anaphylaxis, need rapid investigation
Symptoms – all year or seasonal	Can help to identify the allergen and direct the testing, e.g. respiratory symptoms all year round are unlikely to be due to pollen allergy
Daytime or night-time, indoors or outdoors	Can help to identify the allergen and direct the testing, e.g. symptoms at night are more likely to be due to house dust mites, feathers etc.
Frequent animal contact	Household pets, but also animals of friends and relatives etc.
Does the patient know (or believe) what causes the symptoms?	Often the patient knows what precipitates the symptoms – but they may be wrong
Has exclusion been attempted?	Patients often spontaneously avoid something that makes them unwell

TABLE 30.19 Indications for investigating patients with allergies

Reason	Notes
Severe symptoms	Anaphylactic reaction or symptoms affecting patient's overall health
Progressively worsening symptoms	Every time the patient encounters the allergen, the symptoms appear to be worse – next time it could be anaphylactic
Allergen is difficult to exclude or hidden	Foods, e.g. nuts, eggs, occupational allergens, particular tree pollens, favourite pet
Patient needs a diagnosis	Showing a patient test results may help them accept their condition and accept the reason for the proposed management, particularly if it involves dietary or other restrictions

reasons for investigating patients with allergies are shown in [Table 30.19](#).

There are two general ways of investigating allergy but they are only useful as an adjunct to a good clinical history and examination. They are skin prick testing and measurement of IgE (total and specific). Both are widely used but have advantages and disadvantages, as shown in [Table 30.20](#). In practice, both methods of investigation have a role in the allergy clinic and the allergist will use a combination depending upon a patient's age, symptoms, the putative allergens and the overall clinical picture.

Raised plasma total IgE concentrations are seen in allergic disease, parasitic infections, immune deficiencies, malignancies, liver disease and some viral infections.

With regard to allergies, there are some important caveats. While a raised total IgE concentration suggests a high risk of allergic disease, it does not directly relate to the severity of disease. An IgE concentration within the reference range (particularly in children) does not exclude significant allergy and if there are strong clinical indications, further investigation is warranted.

The IgE that has activity against a particular defined antigen is called specific IgE. The term RAST refers to the technique that was first developed to detect and quantify IgE to specific allergens and, although now scientifically outdated, the term is still commonly used. Specific IgE is reported as k_A U/L or converted to a grade based upon the concentration; both these results can be used as a guide to the patient's potential reactivity to that allergen (or group of allergens). (The subscripted 'A' refers to 'allergen'; however, the units are arbitrary in that they are calibrated against a total IgE standard rather than an allergen-specific IgE standard.) The grades corresponding to various ranges of specific IgE concentrations, and their significance, are shown in [Table 30.21](#). Like total IgE, the specific IgE does not correlate with disease activity between patients. However, in individual patients (particularly children), there is some predictive value in specific IgE concentrations, for example to predict whether a challenge test would be appropriate, so reporting specific IgE in units rather than grades is now accepted.

Anaphylaxis

Anaphylaxis is a medical emergency. It is the sudden, generalized shock and collapse that occurs when patients react to substances to which they are exquisitely sensitive. The clinical features (see [Table 30.22](#)) are induced

TABLE 30.20 Comparison of skin prick testing and specific IgE in the investigation of allergy

Property	Skin prick testing	Specific IgE
Specificity	Atopic patients may show positive skin tests to several antigens, not all of which cause symptoms	Atopic patients may show positive specific IgE to several antigens, not all of which cause symptoms
Applicability	In vivo test, so assesses the patient's response to the allergen	In vitro test, so more a test of the potential to react to an allergen
Use in children	Can be difficult in young children because they have to sit still for ~20 min	Requires a blood test but many allergens can be checked on a single specimen
Dermatological symptoms	Can be impossible in patients with extensive skin disease, e.g. eczema, dermatitis	Dermatological disease does not interfere
Risk	Must be done under medical supervision with resuscitation facilities available	As for a blood test
Convenience	Results available in the clinic within ~20 min	Samples need to be tested in the laboratory; results in days or weeks
Antihistamines	Limited use if the patient is taking antihistamines	Antihistamines do not interfere
Sensitivity	Can be insensitive for diagnosis of food allergy	Can give negative results even with highly suggestive history. Specific IgE may be negative in the first few weeks after a severe reaction
Allergens	Can be difficult if the allergen is toxic or insoluble test material often less purified and similar to native allergen	Can be used for toxic or insoluble allergens Allergens are highly purified and can show variability between different manufacturers
Standardization	Subjective with interoperator variability Saline and histamine must be used as negative and positive quality controls Results reported as wheal and flare diameter	Objective with results reported in arbitrary k_A U/L of IgE (or grades) related to the international reference preparation External QA schemes available to validate results

TABLE 30.21 Guide to the interpretation of specific IgE^a measurements

Units	Grade	Interpretation	Significance (k _A U/L)
<0.35	0	Negative	Significance of grades 1–3 varies depending on allergen
0.35–0.7	1	Weakly positive	Grade 1 to foods or moulds may be significant, but to inhaled allergens is of doubtful significance
0.7–3.5	2	Positive	Positive specific IgE ^a indicates potential to react
3.5–17.5	3	Positive	
17.5–50	4	Strongly positive	
50–100	5	Strongly positive	
>100	6	Strongly positive	

^aHistorically, specific IgE was measured by radioallergosorbent assay (RAST); this method is no longer available but the term RAST is still used as a synonym for a test for specific IgE.

TABLE 30.22 The clinical features of anaphylaxis

Organ affected	Feature
Respiratory system	Upper airway obstruction due to oedema Bronchospasm Oedema Asphyxia
Cardiovascular system	Hypotensive shock due to vascular pooling Cardiovascular collapse
Gastrointestinal system	Vomiting, nausea, diarrhoea, abdominal cramps
Skin	Erythema, pruritus, urticaria, angioedema
General	Warmth, tingling, flushing, feeling of impending doom

via an IgE-mediated mechanism (anaphylaxis), but similar, non-IgE mediated (anaphylactoid) reactions also occur. The precipitating agents are varied, with drugs, bee and wasp venoms, peanuts, latex, fish and eggs being the most common. The mediators involved are shown in [Table 30.9](#). Initial treatment should be aimed at protecting respiratory and cardiovascular function and tissue oxygenation. Adrenaline (epinephrine), antihistamines and steroids are the mainstays of treatment, together with general measures (e.g. oxygen and intravenous fluid). Long-term management must include avoidance of the precipitating allergen (which in practice may be difficult). Patients with a history of anaphylactic reactions should carry a self-injecting adrenaline device (and be instructed in its use) to administer if they inadvertently encounter the relevant allergen. In some situations, for

example sensitivity to bee and wasp venoms, desensitization can be successful.

The investigation of anaphylaxis. It is important to confirm that an apparently anaphylactic reaction is a type I hypersensitivity reaction, and to try to identify the precipitating agent. Serum tryptase is a good marker of mast cell degranulation and can be used for this purpose; the minimum number of samples required is three: at times zero (as soon after the start of the reaction as possible and preferably within 30 min), at 1–2 h post reaction and a third sample 24 h post reaction or during convalescence. Ideally, additional samples at 3, 6 and 12 h should be requested to help interpretation. The peak concentration (>50 µg/L) is reached 1–3 h after the start of the reaction; values gradually return to normal by 24 h. The timing for further investigation will be guided by the clinical situation but may involve closely supervised skin testing and/or total and specific IgE, generally with interpretation by an immunologist. Hereditary angioedema can cause symptoms very similar to those of a type I hypersensitivity reaction, so measurement of C3, C4 and C1 esterase inhibitor concentrations may be indicated.

Autoimmune diseases

The immune system is tolerant to self-antigens or makes only limited responses to them. Sometimes, this tolerance breaks down and the immune system makes a progressive response that attacks the body's own tissue, resulting in autoimmune disease. The factors that trigger such responses are not well understood; many factors are likely to be involved, for example the patient's immunological status, including their HLA types, any pre-existing inflammation and the presence of any infection. The initial damaging stage is usually cell mediated, and biopsies from affected tissue often show monocytic and lymphocytic infiltration. The production of autoantibodies can be part of the pathogenetic process or a result of a response to hidden antigens that are released as a consequence of cellular destruction. The general characteristics of pathogenetic antibodies and non-pathogenetic antibodies are shown in [Table 30.23](#). The presence and the concentration of many of these antibodies are used to help diagnose and, occasionally, to monitor autoimmune diseases.

Autoantibodies can be detected by their reaction against tissue containing the relevant antigens, using subjective methods such as indirect immunofluorescence. Samples found to be positive by these sensitive (but not very specific) methods are then analysed, using more specific and quantitative methods, for example ELISA. Automated ELISA-based 'screening' techniques are increasingly being used, although the validity in a routine setting is questionable.

It is important that detection of these autoantibodies is not considered to be a 'gold standard' test. They are, at best, markers of disease: they have significant limitations and should be used as part of a diagnostic panel rather than the presence of individual autoantibodies being regarded as diagnostic of any particular disease.

TABLE 30.23 Characteristics of pathogenetic and non-pathogenetic antibodies

Characteristic	Pathogenetic antibodies	Non-pathogenetic antibodies
Antigen distribution	Cell surface antigens, e.g. glomerular basement membrane, acetylcholine receptor, thyroid stimulating hormone receptor	Antigens within the cell
Specificity	Antibodies rarely seen in other conditions	Antibodies often seen in the absence of clinical features
Concentration	Related to disease severity	Unrelated to disease severity
	Mirrors disease activity	Unrelated to disease activity

An arbitrary distinction is often made between organ-specific and non-organ-specific autoimmune diseases, but in reality there is a spectrum of involvement ranging from only one or several organ(s) being affected to autoimmune diseases with systemic manifestations.

Autoimmune endocrine diseases

The autoimmune endocrine disorders include diseases of the thyroid, pancreatic islets, adrenal cortex, gonads, pituitary and parathyroids. More than one endocrine organ can be affected, either simultaneously or at different times during the patient's life. Autoimmune disease affecting the parathyroids or pituitary is rare, and is not discussed further in this chapter.

Thyroid. Autoimmune thyroid disease is discussed in detail in Chapter 19. It includes primary myxoedema, Hashimoto thyroiditis and Graves disease. Measurement of thyroid antibodies is of limited use. Approximately 95% of patients with Hashimoto thyroiditis have antibodies to thyroid peroxidase. There is little point in measuring them in patients with obvious hypo- or hyperthyroidism. Measurement is best reserved for patients with equivocal thyroid function test results and when thyroid disease is suspected, but there are unusual clinical findings. Thyroid peroxidase antibody measurements are not required for monitoring patients on thyroid hormone replacement. Thyroid peroxidase antibodies are also found in Graves disease, but are neither of diagnostic value nor useful in monitoring the response to treatment. They are non-pathogenetic, so their concentration does not change in relation to disease activity.

Patients with untreated Graves disease have antibodies to the thyroid stimulating hormone (TSH) receptor molecule, to which they bind, stimulating the gland and

causing hyperthyroidism. These antibodies are pathogenetic, but while the titre of antibodies correlates with thyroid gland hyperfunction, their measurement is not required routinely for diagnosis, nor is it used for monitoring treatment. The antibody is an IgG and capable of crossing the placenta and causing (transient) neonatal hyperthyroidism. A high maternal antibody titre in the third trimester indicates an increased risk of this condition. Less frequently, antibodies to the TSH receptor block the binding of TSH and cause hypothyroidism.

Pancreas. Type 1 diabetes is caused by progressive autoimmune destruction of pancreatic islet cells. Antibodies to islet cells and the enzyme glutamic acid decarboxylase can be found before clinical onset of disease and for a fairly short period early in its course, often disappearing later. There is considerable interest in how these antibodies participate in the immunological damage and whether the immune process may be susceptible to modulation. There is no clear indication for measuring these antibodies in most patients with diabetes, although they can sometimes be useful in patients with atypical clinical or laboratory findings. (More details can be found in Chapter 15.)

Adrenals. Approximately 60% of cases of Addison disease are due to autoimmune destruction of the adrenal cortex. Antibodies can be demonstrated to a number of adrenal cell components and these can show cross-reactivity to steroid-producing cells of the ovary, testis and placenta, potentially leading to gonadal failure.

Autoimmune polyendocrine syndromes (APS). These are a well-recognized group of conditions where two or more endocrine organs are affected. Table 30.24 shows how the conditions may be classified.

Autoimmune diseases of the gut. Pernicious anaemia (see Chapter 27) causes impaired absorption of dietary vitamin B₁₂. Patients develop a macrocytic anaemia and may develop neurological complications. Antibodies to gastric parietal cells and intrinsic factor are detectable in the majority of patients.

Coeliac disease is characterized by destruction of small intestinal villi, leading to malabsorption. It is triggered by an immunological response to the gliadin component of gluten (a protein in wheat and other cereals) but the endogenous antigen is endomysium, the tissue that surrounds the smooth muscle fibres of the gut. The initial response is T cell mediated but gradually antibodies (IgG and/or IgA) are made to gliadin, endomysium and tissue transglutaminase (the important antigen within the endomysium). Detection of antibodies is valuable in screening for the condition and may be sufficient for diagnosis in patients presenting with typical features (see Chapter 12, but the 'gold standard' diagnostic investigation is evidence of villous atrophy on jejunal biopsy. The UK National Institute for Health and Care Excellence (NICE) has issued a guideline on the diagnosis of coeliac disease (see: www.nice.org.uk/nicemedia/pdf/CG86FullGuideline.pdf).

TABLE 30.24 Classification of autoimmune polyendocrine syndromes

Type (syndrome)	Main clinical features	Underlying cause
Autoimmune polyendocrinopathy, candidiasis, ectodermal dysplasia (APECED) Schmidt	Candidiasis, adrenal failure, hypoparathyroidism	Mutations in the autoimmune regulator element (ARE) gene
Thyroid disease with other autoimmune disease but not Addison disease or hypoparathyroidism)	Adrenal failure, autoimmune thyroid disease, type I diabetes	Antibodies to steroid synthesis enzymes (e.g. 21 α -hydroxylase)
Undefined	Type I diabetes, pernicious anaemia, vitiligo, alopecia	Unknown
	Two or more organ specific autoimmune diseases that do not fall into categories 1–3	

Autoimmune liver diseases. The majority of cases of chronic hepatitis are due to viral infections (particularly with hepatitis B and C viruses) but approximately 20% of cases are autoimmune in origin. The distinction between autoimmune and infectious hepatitis is made mainly on the basis of the clinical features and the results of virological studies. It is important to identify the cause of chronic hepatitis, because interferon treatment (which is used for viral hepatitis) may worsen autoimmune liver disease and immune suppression (used for autoimmune liver disease) can worsen the viral disease. Unfortunately, patients with viral hepatitis sometimes make the very antibodies used in the investigation of autoimmune hepatitis.

Autoimmune hepatitis generally affects young females and is classified into types I and II. The typical age of presentation for type I is in the teenage years (~10–20 years of age), while the peak age of presentation

for type II is 7–8 years. Table 30.25 shows typical results of autoantibody measurements in autoimmune liver diseases. The antibodies are usually detected by indirect immunofluorescence. Considerable skill may be needed to interpret the immunofluorescence patterns and follow-up investigations, for example by ELISA, may sometimes be necessary. In general, high antibody titres are more closely associated with disease than low antibody titres. The hepatic autoantigens are often enzymes; for example, 2-oxoacid dehydrogenase is the specific antigen for the M2 subtype of mitochondrial antibodies that is the most specific for primary biliary cirrhosis.

Liver disease is not usually an indication for measurement of immunoglobulin concentrations, although abnormal immunoglobulin concentrations are common in liver disease, particularly when chronic. For example, patients with primary biliary cirrhosis usually have

TABLE 30.25 Investigations in liver disease, focusing particularly on the measurement of autoantibodies

Cause	Disease	Relevant finding
Infection	Hepatitis	Virus or antibody
Autoimmunity	Autoimmune hepatitis type I	Polyclonally raised IgG and IgA Smooth muscle antibodies Antinuclear antibodies Anti-soluble liver antigen Anti-liver kidney microsomal antibodies ^a
	Autoimmune hepatitis type II	Anti-liver kidney microsomal antibodies ^a Anti-liver cytosol antibodies Anti-soluble liver antigen Immunoglobulins (polyclonal raised IgG)
	Primary biliary cirrhosis	Antimitochondrial antibody ^b Immunoglobulins (predominantly polyclonal raised IgM)
Alcohol	Hepatitis and cirrhosis	Immunoglobulins (predominantly polyclonal raised serum IgA)
Drugs and toxins	Paracetamol, industrial toxins, e.g. carbon tetrachloride, polyvinyl chloride	
Genetic α_1 -antitrypsin deficiency	Prolonged jaundice in neonate	Total α_1 -antitrypsin and phenotype
Wilson disease	Varied but include fulminant hepatic failure, or cirrhosis	Urinary copper, serum copper and caeruloplasmin
Malignancy	Primary liver cancer and metastases (e.g. from bowel primary)	Tumour markers, e.g. α -fetoprotein, carcinoembryonic antigen
Obesity	Fatty liver disease	

^aLiver kidney microsomal antibodies occur only in a small number of patients with autoimmune hepatitis.

^bAntimitochondrial antibodies are virtually pathognomonic of primary biliary cirrhosis.

markedly raised polyclonal IgM concentrations; patients with alcoholic liver disease tend to have raised polyclonal IgA concentrations, and patients with persistent infections or inflammatory responses show raised polyclonal IgG and IgA concentrations. However, these patterns are non-specific and of negligible diagnostic value. Serial measurements of IgG may, however, be of value in monitoring patients with autoimmune chronic hepatitis.

Autoimmune skin diseases. The blistering skin diseases, pemphigus and pemphigoid, are caused by antibodies to components of the skin. These IgG antibodies bind to antigens in the skin, activate complement and disrupt the skin basement membrane (pemphigoid) or skin intercellular cement (pemphigus) causing intraepidermal blisters.

Autoimmune kidney diseases. Immunological damage due to the deposition on the glomerular basement membrane (GBM) of immune complexes (either to antigens of the GBM (as in Goodpasture syndrome) or formed in the circulation), is an important pathogenetic mechanism in renal disease. When the clinical features suggest a diagnosis of autoimmune kidney disease, evidence from renal biopsy often provides a definitive diagnosis. However, laboratory data can reduce the need for biopsy. The antibodies particularly implicated in autoimmune kidney disease are shown in Table 30.26. Laboratories should have protocols that have been agreed with the local renal physicians for investigating patients with renal disease, but detection of antinuclear antibody (ANA), antineutrophil cytoplasmic antibodies (ANCA) and GBM antibodies should ideally be available on an urgent basis (although, note that in immunological terms, this means results being available within 24 h of request). Patients with ANCA-associated vasculitis and GBM disease can also have pulmonary involvement (Goodpasture syndrome can manifest as haemoptysis). The vascularity of the lungs and kidneys makes them particular targets for vasculitic disease and the antibodies formed to either glomerular or alveolar basement membrane often show cross-reactivity with the other membrane. The concentrations of GBM antibodies and of antibodies to proteinase 3 (and myeloperoxidase) can be used to monitor patients' responses

to treatment and, in the vasculitides, to provide an early warning of exacerbations of disease.

Testing for renal autoantibodies is essential early in the investigation of unexplained renal impairment. Such patients should also have their serum and urine investigated for paraproteins (including Bence Jones proteinuria). The investigation of samples for cryoproteins is often forgotten but can be a vital test. The cryoproteins of relevance are not those that are seen with large monoclonal components that precipitate in the cold in vivo. Rather, they are high molecular weight immune complexes associated with infectious agents, for example hepatitis B and C viruses, which lodge in blood vessels, activating complement and causing vasculitis. This process can affect any organ, but the kidneys, with their large vascular network, are a particular target. This cryoprecipitation is a laboratory artefact, generated by storage of serum samples in the cold; the amount of precipitate can be very small, but careful isolation and investigation typically reveals monoclonal or polyclonal IgM showing rheumatoid factor activity.

Autoimmune articular diseases. These diseases are also referred to as connective tissue diseases. They are discussed in more detail in Chapter 32. There is considerable overlap between the clinical features seen in individual conditions. Autoantibodies are used as markers of the diseases; their presence contributes to, but does not establish, the diagnosis.

Rheumatoid arthritis (RA). Rheumatoid factor (RF) is an autoantibody (usually IgM) directed against the Fc portion of IgG. It is not specific for rheumatoid arthritis, also occurring in infections, malignancies, persistent inflammatory conditions and in up to 10% of healthy adults, with increasing frequency in elderly patients. A raised RF is one of the diagnostic criteria for RA but is not essential for the diagnosis. There is a correlation between higher RF concentrations and more severe disease and poorer long-term prognosis in RA; less severe disease is seen in seronegative patients. Rheumatoid factor is not of value in monitoring RA: measurement of C-reactive protein (CRP), reflecting the severity of inflammation, is superior. Changes in CRP precede both radiological and clinical changes.

TABLE 30.26 Antibody mediated renal disease

Disorder	Serological findings	Biopsy findings
Antiglomerular basement disease (Goodpasture syndrome)	Serum antibodies that react with normal basement membrane (GBM ^a)	Linear IgG, C3 along basement membrane
SLE nephritis	Antinuclear antibodies ^a (see text)	Granular IgG, IgM, C3, C4 along GBM
Vasculitis, e.g. Wegener granulomatosis	Anti-neutrophil cytoplasmic antibodies ^a (ANCA) Staining in the cytoplasm hence c-ANCA and is associated with antibodies to proteinase 3	Inflammatory – nothing specific
Microscopic polyarteritis	ANCA Staining in the perinuclear area hence p-ANCA and is associated with antibodies to myeloperoxidase	
Cryoglobulinaemia	Cryoglobulins (proteins that form a precipitate on cooling of serum)	IgG, IgM, C3 along capillary wall

^aThe indirect immunofluorescent staining patterns are shown in Figure 30.12.

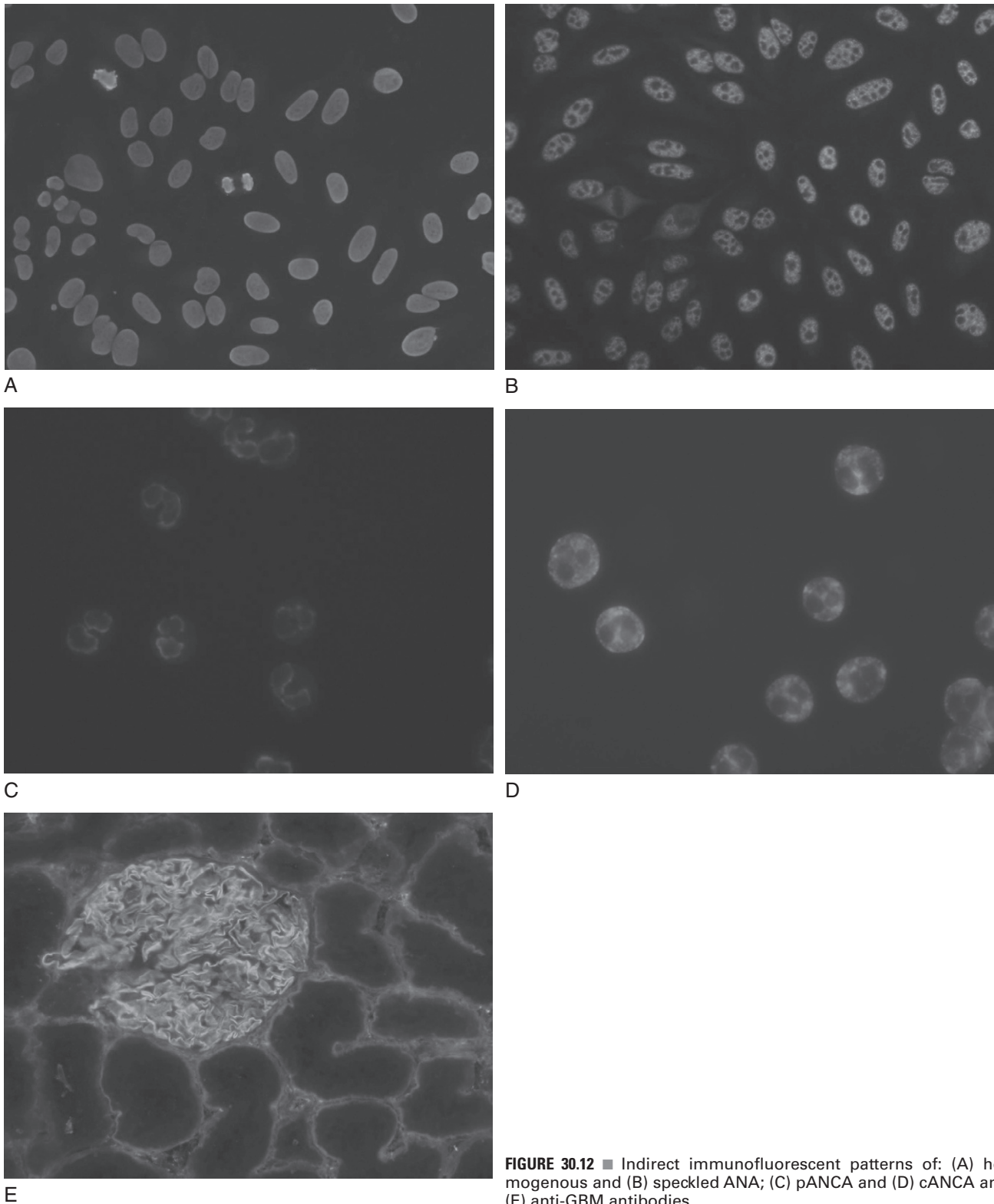


FIGURE 30.12 ■ Indirect immunofluorescent patterns of: (A) homogenous and (B) speckled ANA; (C) pANCA and (D) cANCA and (E) anti-GBM antibodies.

The measurement of IgG antibodies to citrullinated peptides (anti-cyclic citrullinated peptide, CCP) is increasingly being used in investigation of patients with RA. It shows similar clinical specificity and sensitivity to RF but the presence of CCP antibodies can precede overt signs of RA. A combination of RF and

CCP antibodies is becoming important to stratify patients with RA and to direct treatment, e.g. use of anti TNF. The UK National Institute for Health and Care Excellence has issued a guideline on the management of RA in adults (see: <http://www.nice.org.uk/guidance/cg79>).

Other connective tissue diseases. The detection of antibodies to nuclear components is one of the most important laboratory tests in the investigation of patients with suspected systemic lupus erythematosus (SLE), Sjögren syndrome and scleroderma. The term ‘antinuclear antibodies’ is used to describe this diverse group of autoantibodies that react with nuclear and cytoplasmic antigens that are common to all nucleated cells. The antigens typically have functions in the cell cycle, or in transcription and translation.

Antinuclear antibodies. Antinuclear patterns are detected by indirect immunofluorescence using either frozen sections of animal tissue or cultured human cancer cell lines, most commonly the epithelial cell line HEp-2. This is usually a very sensitive test and specimens that test positive for antinuclear antibodies then need testing for one of the nuclear antigens by specific methods. These specific tests use purified or partially purified antigens. There is a move towards using combined specific tests instead of the subjective immunofluorescence assays, but these methods are not yet widely accepted as potential replacements.

The pattern of the immunofluorescence reaction gives some indication of the antigen specificity because of the different distribution of antigens within the nucleus and cytoplasm. Figure 30.13 shows a diagram of a nucleus with the distribution of the important antigens. The more frequently identified and clinically useful antibodies are listed in Table 30.27. Specificities other than those listed rarely have clinical significance, are usually present at low concentrations and can be found non-specifically in many infectious and inflammatory conditions.

Antibodies to double-stranded DNA. Antibodies to single-stranded (ss) DNA are frequently present in SLE but can also be demonstrated in many other inflammatory conditions and in infections. Antibodies to double-stranded (ds) DNA are one of the important diagnostic criteria in SLE. The most frequently used methods for the quantification of anti-dsDNA are ELISAs, but the immunofluorescence assay using *Critidia luciliae* is usually a more specific (although less sensitive) assay. High antibody concentrations to

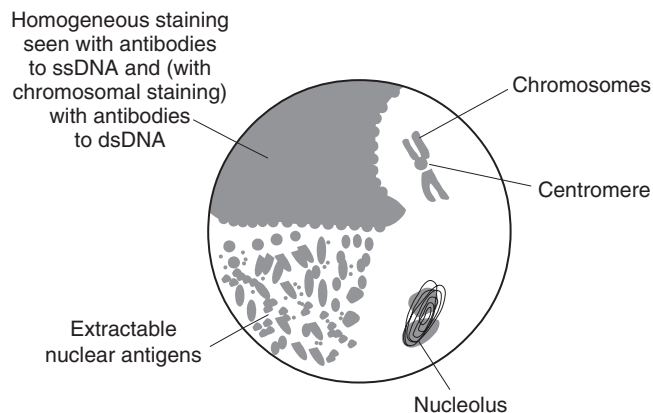


FIGURE 30.13 ■ Diagram of a nucleus showing staining patterns; ss, single-stranded; ds, double-stranded.

TABLE 30.27 Antinuclear antibody patterns and their disease-associated antigens with reported frequency of detection (%)

ANA pattern	Antigen	Disease association
Homogeneous	DNA (double and single stranded)	SLE (60%)
	Histone	Drug-induced lupus (95%) SLE (60%)
	Topoisomerase-1 (Scl-70)	Progressive systemic sclerosis (15–70%)
Speckled	Sm RNP	SLE (20%) SLE-Sjögren overlap (100%) SLE (25%)
	SSA/Ro	Sjögren syndrome (60%) SLE (35%)
	SSB/La	Sjögren syndrome (40%) SLE (15%)
Centromere	CENP	Limited cutaneous scleroderma (formerly CREST ^a syndrome) (7–21%)
Nucleolar	PM-Scl	Polymyositis (8%) Myositis/scleroderma overlap (50%)
	Nucleolar RNA	Scleroderma (5–43%)
Cytoplasmic	Histidyl-tRNA synthetase (Jo-1)	Polymyositis (2%)

^aCREST: calcification, Raynaud phenomenon, (o)esophageal involvement, sclerodactyly and telangiectasia.

dsDNA are associated with renal complications of SLE and quantification of dsDNA antibodies is useful (together with C3 and C4 concentrations) to monitor the disease and response to treatment.

Antibodies to extractable nuclear antigens. Numerous antibodies to various nuclear or cytoplasmic protein antigens (commonly termed the extractable nuclear antigens or ENAs) are detectable in connective tissue disorders (see Table 30.27). They may be of diagnostic value but, in contrast to antibodies to dsDNA, there is negligible value in using these tests for monitoring disease activity.

Antinuclear antibodies in pregnancy. Antinuclear antibodies are of the IgG class and can therefore be transported across the placenta to the fetus. Antibodies to dsDNA can cause a transient, self-limiting, neonatal lupus syndrome. Antibodies to the Ro antigen cross-react with the fetal heart tissue and can cause neonatal heart block.

Antiphospholipid antibodies. These are a group of antibodies that are associated with an increased tendency for thrombosis. The most clinically relevant are

antibodies to cardiolipin, β_2 -glycoprotein 1 and 'lupus anticoagulant'. They may occur in some patients with SLE but can also be found in patients with recurrent miscarriages, unexplained transient ischaemic attacks, strokes or thrombotic events.

LYMPHOID MALIGNANCIES

Lymphocytes go through a complex process from the start of their lives as naïve lymphocytes to end-differentiated, functional lymphocytes. The complexity of the process means that there are many opportunities for a cell to undergo a malignant transformation that may result in malignant disease. The classification of lymphoid malignancies is based upon the normal physiological pathways and the place in this pathway where the malignant transformation has occurred. The leukaemias are malignancies of cells whose mature forms are normally found in the blood, while the lymphomas are tumours of the non-recirculating lymphoid cells. The diagnosis and management of lymphoid malignancy is predominantly the remit of haematologists. Flow cytometric analysis of peripheral blood for panels of CD markers is important as are cytological examinations of either lymph node or bone marrow biopsies. The detection of the products of monoclonal B cells, monoclonal proteins, is, however, a vital test in the investigation of patients with suspected B cell malignancy (the category that includes myeloma). There are no specific biochemical investigations for the diagnosis and management of malignancies of the T cell lineage, and the subsequent discussion therefore focuses on malignancies of B cells.

B lymphocytes and monoclonal proteins

Early B lymphocytes, with their randomly rearranged B cell receptors (IgM), leave the bone marrow and traffic around the lymphoid tissue, searching for appropriate antigens. Once a lymphocyte meets an antigen, there is a process of maturation to secrete high-affinity antibody (after class switching) and to maintain a memory response. Plasma cells are end-differentiated B cells that reside in the bone marrow and secrete immunoglobulins. There is normally considerable microheterogeneity in the immunoglobulin molecules. This results in polyclonal immunoglobulins, representing all classes and light chain types, with a large number of different idiotypes produced by a large number of different B cell clones. A monoclonal immunoglobulin of a single heavy chain class, light chain type and idio type, results from the proliferation of a single clone of B cells. Monoclonal B cells, especially those that have undergone malignant transformation, do not necessarily produce normal, intact immunoglobulin: in approximately 20% of patients with myeloma, they produce only monoclonal free light chains (detectable in the urine as Bence Jones protein, BJP). Almost all variations on the normal immunoglobulin molecules are possible: monoclonal monomeric (7S), instead of the normal pentameric IgM, may be secreted, monoclonal fragments, for example free heavy chains,

half molecules and truncated molecules can all occur; so, too, can combinations of immunoglobulin fragments and intact monoclonal immunoglobulins. The class of immunoglobulin produced by the tumour can give a clue to the place in the B cell development where the malignant transformation occurred. For example, tumours of early B cells often secrete monoclonal IgM, while tumours of end-differentiated B cells tend to secrete IgG or IgA.

Monoclonal proteins or fragments appear as homogeneous, compact bands on electrophoretic separations. These bands should be called monoclonal proteins but are often called paraproteins, M components or M proteins. Monoclonal proteins detectable in serum are most often intact immunoglobulins; if there is renal damage these proteins can leak into the urine. Low molecular weight fragments and monoclonal free light chains (BJPs) can pass readily through normal glomeruli and appear in the urine in the absence of renal damage but may also be detectable in the serum.

Clinical significance of monoclonal proteins

The finding of a monoclonal protein in the serum and/or urine can be associated with both malignant and benign conditions, as indicated in [Table 30.28](#).

In malignant diseases, paraproteins are sensitive tumour markers (see below), but their presence in (apparently) non-malignant diseases limits their specificity. However, immunoglobulin fragments are rarely found in non-malignant monoclonal B cell expansions and their presence is thus highly suggestive of malignant B cell proliferation. The clinical features associated with paraproteinaemia can be due to the effects of the underlying tumour or to the properties of the monoclonal protein. Features related specifically to paraproteins are listed in [Table 30.29](#). Some paraproteins also show autoantibody activity, for example against other

TABLE 30.28 Conditions associated with the production of monoclonal proteins

Category	Condition
Malignant	Multiple myeloma, plasmacytoma, Waldenström macroglobulinaemia, lymphoma (of B cell origin), chronic lymphocytic leukaemia
Non-malignant (asymptomatic)	Benign Transient Monoclonal gammopathy of unknown significance (MGUS)
Non-malignant (symptomatic)	Amyloidosis Cryoglobulinaemia and immune complex disease Primary cold agglutinin disease

There are also dermatological and neurological diseases where the incidence of paraproteinaemia is greater than can be accounted for by coincidence. Amyloidosis with neurological and dermatological manifestations can occur as a complication of malignant diseases.

TABLE 30.29 Clinical manifestations of the physicochemical properties of paraproteins

Condition	% of patients with paraproteins	Main Ig classes
Hyperviscosity syndrome	8	IgG3, IgA and, particularly, IgM
Amyloid deposition	8	Bence Jones
Renal tubular obstruction	6	Bence Jones
Thromboembolism	5	IgG
Immune complex disease	4	IgM, IgG
Monoclonal cryoglobulinaemia	2	IgM, IgG
Fanconi syndrome	1	Bence Jones

immunoglobulins (e.g. rheumatoid factor) in immune complex disease and cryoglobulinaemia, against the I red cell antigen in primary cold agglutinin disease, and against myelin-associated glycoprotein in peripheral neuropathy.

Monoclonal proteins can also occur secondarily to infection and as part of immune reconstitution post-transplantation; in these situations, they are generally transient, disappearing over weeks or months. Monoclonal proteins may also be found in patients without any overt clinical signs of malignancy; this is particularly true in the elderly.

Prevalence of monoclonal proteins

The prevalence of monoclonal proteins is dependent on the population being considered. In symptomatic populations (e.g. hospital patients), paraproteins are predominantly associated with B cell malignancy, of which myeloma is the most common. In asymptomatic populations, the prevalence depends on the sensitivity of the electrophoresis used to detect the paraprotein and the characteristics of the group that is screened. The incidence of paraproteins rises with age; a survey of over 9000 apparently healthy individuals showed paraproteins in 0.1% of those aged 20–30 years, in 0.6% by age 60 years and in 6% of subjects aged 90–99 years.

Laboratory investigation of paraproteins

The detection, typing and quantification of monoclonal components in serum and urine are essential laboratory investigations in the diagnosis and monitoring of both malignant and non-malignant monoclonal B cell expansions. A serum paraprotein and/or urine BJP is a highly sensitive tumour marker for certain B cell malignancies, particularly myeloma (>95%) and Waldenström macroglobulinaemia (100%). In these conditions, the monoclonal component is indeed a diagnostic feature. However, they are not entirely specific for these conditions and can be found in other B cell malignancies and non-malignant diseases (see [Table 30.28](#)). It is important to note that the detection of a monoclonal protein does not necessarily

BOX 30.2 Findings that should prompt the investigation of serum and urine for a monoclonal immunoglobulin

Symptoms

- Backache (especially localized or with local tenderness)
- Lassitude
- Recurrent infection

Clinical syndromes

- Renal impairment
- Nephrotic syndrome
- Peripheral neuropathy
- Hyperviscosity syndrome
- Carpal tunnel syndrome
- Malabsorption

Radiological findings

- Osteolytic lesions
- Pathological fractures

Haematological findings

- Normochromic, normocytic anaemia
- Raised erythrocyte sedimentation rate (ESR)

Biochemical findings

- Raised serum total protein
- Raised serum globulin (total protein minus albumin)
- Proteinuria
- Abnormal results of renal function tests
- Subnormal immunoglobulin concentrations
- Hypercalcaemia

indicate that the patient has myeloma or, indeed, any malignant transformation of B cells, nor does the absence of a paraprotein completely exclude the presence of myeloma or B cell malignancy.

The presence of a monoclonal immunoglobulin may be suspected from the patient's clinical presentation or from results of other investigations. The findings that should, in the absence of another explanation, prompt investigation for paraproteinaemia, are shown in [Box 30.2](#).

Identification of paraproteins

The detection of monoclonal proteins is one of the two clear indications for the measurement of serum IgG, IgA and IgM concentrations and protein electrophoresis; these investigations should be done as a logical group of tests for optimum interpretation; this approach is in line with the consensus recommendations of the International Myeloma Workshop Consensus Panel (see [Further reading](#), below). Protein electrophoresis is the only reliable method for the detection of paraproteins in serum and urine. Abnormal bands (in the serum), once detected by electrophoresis, can be quantified by scanning densitometry or directly from the capillary zone electrophoresis trace. Immunofixation should be used to type the monoclonal band; this will confirm clonality and identify the heavy and light chains.

Automated capillary zone electrophoresis (CZE) systems are becoming the most common method for serum

protein electrophoresis, although agarose gel based methods are still in frequent use. Capillary zone electrophoresis techniques for urine protein electrophoresis are available but sample preparation is complex and the methods not sufficiently robust for routine practice. Serum proteins are separated by electrophoresis at pH 8.6; both agarose and CZE systems generate five major zones. Figure 30.14 shows the typical mobility of the major serum proteins. Agarose separations should be long enough (3–4 cm) to allow good separation of zones and show a good spread of the β - γ zone, which is achieved by properties of the agarose gel and buffer system that produce high endosmotic flow. The majority of serum paraproteins migrate in the β or γ zones, although they may be found anywhere on the electrophoretic separation between the α_1 zone and the post- γ region. Monoclonal heavy chains may show as diffuse zones running very anodally in the α_2 - β region of the electrophoretic separation; IgD paraproteins can also form a diffuse zone in the β - γ region. Representative examples of serum protein electrophoretic patterns are shown in Figure 30.14.

Paraproteins may be missed if they are present at low serum concentrations (<5.0 g/L), when their mobility coincides with other bands such as β -globulins, or where there is no suppression of normal background immunoglobulin concentrations. This is most often seen with low-concentration IgM or IgA paraproteins. Therefore, whenever there is a raised concentration of IgA or IgM with no clear increased staining in the β - γ region that would indicate a polyclonal increase, immunofixation should be done, as it should be also if there is a high clinical index of suspicion of B cell malignancy but an apparently normal electrophoresis. IgD paraproteins and free heavy chains are susceptible to post-synthetic degradation, which results in diffuse paraprotein bands on electrophoresis, and these may be missed if present at low concentrations or if the operator is expecting to see a narrow band rather than a diffuse zone. Precipitation of paraproteins on cooling of blood samples (monoclonal cryoglobulins) and subsequent removal with the clot, can also result in failure to detect a significant paraprotein. There are a number of proteins that can be mistaken for monoclonal proteins in electrophoresis separations.

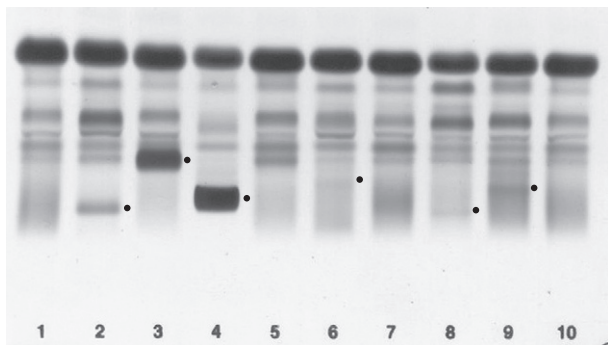


FIGURE 30.14 ■ Representative serum electrophoresis patterns. Tracks 1 and 10 are dilutions of normal serum. Dots to the right of some of the tracks indicate bands that require immunofixation for further investigation.

These include:

- allotypic variants, e.g. of α_1 -antitrypsin, may result in two bands in a normally homogeneous region
- a haptoglobin–haemoglobin complex as a result of in vitro haemolysis, causing splitting of the normally homogeneous α_2 zone
- acute phase proteins, e.g. C-reactive protein, increases, which may result in additional bands being present
- fibrinogen in plasma or inadequately clotted samples, which results in an additional band in the fast γ region
- lipoproteins, which may give distinct bands in the β region in agarose systems
- haemoglobin, which migrates in the β region.

Urine should be examined in every patient in whom B cell malignancy is being considered. This is especially important if there is suppression of the serum immunoglobulins or if there is renal impairment, and it should be done even where no serum monoclonal component is detected. It is also important to monitor the amount of BJP during a patient's follow-up, even when BJP is not detected at presentation, because it can appear later in the course of the disease ('Bence Jones escape').

The presence of BJP provides strong evidence of malignancy, although it can occur in apparently benign conditions. The International Federation for Clinical Chemistry (IFCC) recommends a sensitivity for the detection of BJP of 10 mg/L: high-sensitivity electrophoresis of urine is therefore essential. This can be achieved by using very sensitive stains when electrophoresis is performed on unconcentrated urines, or by concentrating urines at least 100-fold. The concentration devices separate the water, ions and low molecular weight peptides from higher molecular weight constituents across a semipermeable membrane. The molecular exclusion of these membranes should be between 11.5 kDa and 13.0 kDa. Some concentration systems are static; others require centrifugation to achieve separation. Urine concentrates should then be analysed alongside serum in the agarose systems. Ideally, a trace of albumin should be visible in every urine sample; failure to detect any albumin on the electrophoretic separation should prompt further concentration and re-analysis. The renal damage associated with Bence Jones proteinuria often results in complex patterns; immunofixation improves both specificity and sensitivity, so is very useful to help resolve these patterns. Representative examples of urine protein electrophoretic patterns are shown in Figure 30.15. Low concentrations of BJP can be seen with a significant glomerular proteinuria in patients with light chain renal amyloidosis. Any urine with a marked glomerular proteinuria should be investigated by immunofixation, even in the absence of a band suggestive of BJP.

Several proteins can be mistaken for monoclonal protein in urine electrophoresis separations, particularly where there is an element of tubular proteinuria. They include the following:

- α_1 -microglobulins
- lysozyme (migrating in the slow γ region)
- degraded fragments of protein of glomerular origin

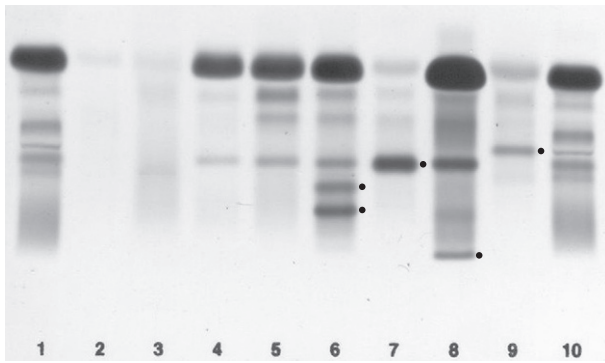


FIGURE 30.15 ■ Representative urine electrophoresis patterns. Tracks 1 and 10 are dilutions of normal *serum*; all urine samples contain a trace of albumin. Significant bands that require immunofixation for further investigation are indicated by dots to the right of the respective tracks.

- seminal fluid proteins (very rare)
- β_2 -microglobulin (can produce a prominent band when present in high concentrations).

Measurement of serum free light chains is now routinely available. A combination of serum protein electrophoresis and free light chain measurement has been suggested as a suitable screening protocol for B cell malignancy. This approach is now being used in some laboratories and is further discussed on p. 596.

Typing of monoclonal immunoglobulins in serum and urine

Immunofixation. Once a paraprotein band has been detected, it is important to identify the heavy and light chain components. The major indications for immunofixation are to:

- confirm clonality of a band detected by electrophoresis
- test for α , γ or μ heavy chains and κ and λ light chains (routinely)
- test for δ and ϵ heavy chains when a serum shows monoclonal light chains without a corresponding α , γ or μ heavy chain
- exclude monoclonal components being present at low concentrations, even where no band is apparent

on electrophoresis but when there are appropriate clinical indications, e.g. amyloid

- exclude the presence of monoclonal IGA or IGM if there are raised concentrations without β - γ fusion (the appearance when these bands lose their separate identities)
- investigate the possibility of an apparent paraprotein in serum or urine being caused by a high concentration of another protein (e.g. fibrinogen, complement components, β_2 -microglobulin) (using antisera specific to these proteins)
- detect minimal residual disease or complete remission following haematopoietic stem cell transplantation for myeloma when no monoclonal component is seen on the electrophoretic separation.

When monoclonality has been confirmed, determination of the paraprotein type may provide useful additional information about the underlying tumour and prognosis. The distribution of heavy chain classes in B cell dyscrasias is shown in Table 30.30. The light chain distribution of the common paraproteins is similar to that of normal polyclonal immunoglobulins, with a predominance of κ over γ (about 2:1). IgD paraproteins are usually of γ light chain type, although κ may be found, and the few IgE paraproteins that have been reported have all been of κ light chain type.

The electrophoresis pattern may change during the course of a patient's disease or treatment. For example, some patients may start producing BJP after having not done so at the time of diagnosis. Complete disappearance of a paraprotein is rare but is occurring increasingly with treatment regimens using high-dose chemotherapy and bone marrow transplantation. An oligoclonal banding pattern is often seen in patients after stem cell transplantation and it is important to distinguish this from the original paraproteinaemia.

Immunofixation enhances the sensitivity of gel electrophoresis both by removing background staining and selectively increasing the amount of protein (by adding antibodies) in the band of interest. Representative examples of serum and urine immunofixation patterns are shown in Figures 30.16 and 30.17, respectively. The choice of antisera is paramount: problems with this technique are often related to the use of poor antisera. This is

TABLE 30.30 Paraprotein types in B cell dyscrasias

Monoclonal type	All paraproteins (%)	Myeloma (%) ^a	Waldenström macroglobulinaemia (%)	Other B cell diseases (%) (NHL/CLL ^b)	MGUS (%)
IgG	53	53 ^c (40)	–	40	60
IgA	22	22 ^c (22)	–	10	35
IgM	11.4	0.5 (19)	100 ^e	50	5
IgD	1.3	1.51 ^{c,d} (14)	–	–	–
IgE	<0.001 ^e	0.1 (30)	–	–	–
BJP only	11.6	21 (5.4g/24h)	–	–	(+)
Biclonal	1.5 ^f	Unknown	–	–	–
Non-secretory	–	1	–	–	–

^aValues in brackets = typical serum concentrations in g/L (except BJP); ^bNHL, non-Hodgkin lymphoma; CLL, chronic lymphatic leukaemia; both are tumours of B cell origin. ^cCan also show Bence Jones protein; ^dusually γ light chain; ^eusually κ light chain; ^fdoes not include intact monoclonal Ig with BJP or multiple bands of the same paraprotein type.

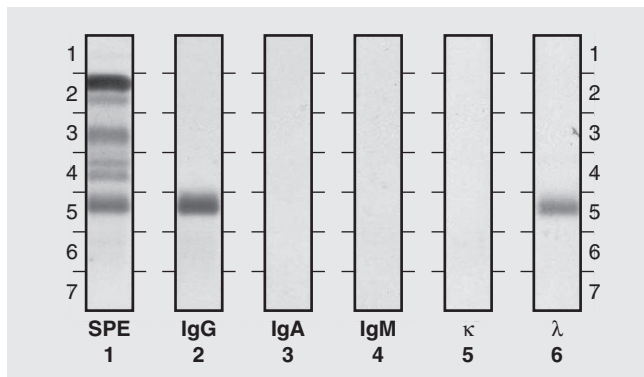


FIGURE 30.16 ■ Representative immunofixation of a serum sample containing an IgGλ paraprotein.

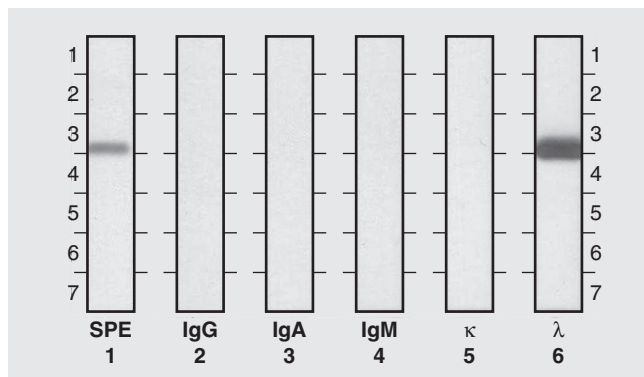


FIGURE 30.17 ■ Representative immunofixation of a urine specimen containing a λ Bence Jones protein.

especially true for the antisera to free κ and λ light chains. Such antisera are available but often react poorly. The antisera most widely used for immunofixation react with both free and bound immunoglobulin heavy and light chains. The presence of free light chains can be inferred if no heavy chain reaction is present. Immunofixation is not a quantitative method; antisera often show greater binding to free than to bound light chains.

A major problem with immunofixation of urine is distinguishing between monoclonal light chains and light chain fragments generated from normal immunoglobulin catabolism, which result in a 'ladder-like' pattern, particularly in the immunofixation reaction with antiserum to κ light chains. This 'κ banding' pattern is most often seen in elderly patients with inflammatory conditions. Examples of BJP and κ banding are shown in Figures 30.17 and 30.18.

Immunofixation techniques are not applicable in automated CZE systems; therefore, the technique of immunosubtraction has been developed for these systems. Immunosubtraction techniques work well for large monoclonal proteins but can lack sensitivity for small paraproteins and are inflexible; for example no reagents are available for typing IgD or IgE paraproteins or other non-immunoglobulin proteins.

In rare circumstances, the presence of free heavy chains may need to be confirmed. This is done by an adaptation

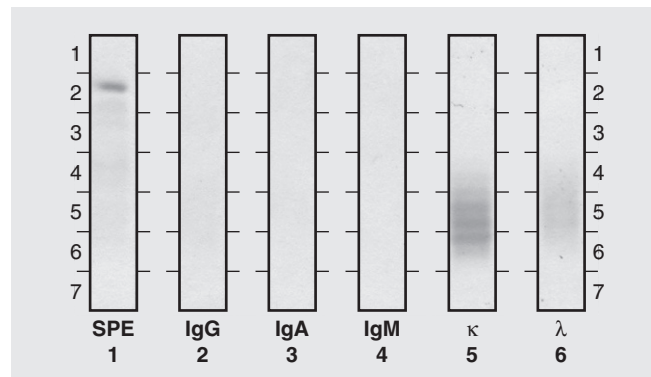


FIGURE 30.18 ■ Representative immunofixation of a urine specimen showing κ banding (lane 5) (immunofixation will vary between manufacturers, in particular, the extent of separation).

of gel electrophoresis called immunoselection, or by molecular weight and immunoblotting methods that are best referred to expert laboratories.

Quantitation of monoclonal components

The concentration of a monoclonal component does not, in general, predict tumour mass. However, in individual patients, concentration does reflect tumour mass and is therefore an essential component in both the initial investigation and the assessment of the response to treatment of patients with B cell malignancies. Table 30.30 shows the average paraprotein concentration at presentation with myeloma with respect to the paraprotein type.

The only reliable method of paraprotein quantitation is to measure the percentage of the monoclonal protein (by scanning densitometry or from CZE read-out), with respect to the total electrophoretic separation or to the globulin fraction. The stains used in protein electrophoresis show differential dye binding between albumin and globulin; the more precise estimation of paraprotein is therefore derived from the percentage of relative dye binding of the paraprotein band compared with the total globulin fraction rather than the total protein. The CZE measuring system uses a UV detector system that overcomes the differential dye binding and also limits problems of non-linear dye binding at high paraprotein concentrations. Serum total protein is most often measured by biuret-based methods. A useful check is to take the sum of the paraprotein and albumin concentrations and add 10–15 g/L (to account for the other serum proteins): the result should be approximately equal to the measured serum total protein.

It is possible to quantitate reliably a serum paraprotein down to a concentration of 1 g/L, provided there is suppression of the background immunoglobulins. Monoclonal proteins with electrophoretic mobility in the β region can be difficult to measure. However, the concentration of the β-globulins is relatively constant. Therefore, when a paraprotein band migrates with β-globulins, the total band can be measured: significant changes in that concentration are likely to be due to changes only in the concentration of the monoclonal

component. When it is possible to delineate a monoclonal peak for the purposes of measurement, densitometric quantitation should be reported; however, once the concentration has fallen to an extent that the peak cannot be clearly seen, immunochemical quantitation should be used. The concentration of monoclonal components can reach values in excess of 140 g/L, although this is rare. In this situation, falsely low concentrations can be seen because the paraprotein concentration has exceeded the capacity of the detection system; the monoclonal peak on the CZE tracing may appear with a flat or split top. The frequency of monitoring the concentration of monoclonal proteins will depend upon a patient's clinical condition and treatment regimen. Cumulative reporting of paraprotein studies is invaluable in monitoring patients with myeloma and other paraproteinaemias.

Quantitation of BJP in urine is generally more difficult than quantitation of monoclonal proteins in serum. Urine volume can be very variable, being influenced by factors such as fluid intake, renal function, state of hydration and time of day. Quantitation of BJP should start with the densitometric scan of the stained electrophoretic separation in a similar way to that used for serum paraproteins. The band can be expressed as a percentage of the total protein, but this can give misleading information, especially if the general background proteinuria is increased, making the percentage of BJP relatively lower. This percentage BJP can be used with the urine albumin or creatinine concentration to give the quantity of BJP expressed per gram of albumin or millimole of creatinine. It can also be used with the total protein concentration to generate a concentration in g/L and this result can be used with the 24h urine volume to generate a 24h excretion. There are both practical and analytical issues, with no one method being ideal for all situations. Most importantly, a consistent approach should be adopted so that it is reliable to compare results on the same patient during the course of the disease.

Cryoproteins

Cryoproteins are proteins that aggregate on cooling of the serum or plasma and usually redissolve on warming to 37 °C; immunoglobulins and fibrinogen can all show this property.

Cryofibrinogen is an unusual finding resulting from misdirected synthesis or by post-synthetic modifications of fibrinogen. It can also be (artefactually) induced by heparin; thus collection into EDTA, citrate or oxalate is essential. Cryofibrinogen will only be visible in plasma samples while cryoprecipitating immunoglobulin should be visible in both serum and plasma stored at 4 °C. Cryofibrinogenaemia typically produces skin lesions or thrombophlebitis migrans and patients are managed by dermatologists or haematologists.

Cryoprecipitating immunoglobulins are classified into types I, II and III, based on whether they are mono- or polyclonal, their approximate concentrations and other properties; details of this classification are shown in [Box 30.3](#). Individual cryoproteins show particular

BOX 30.3 Classification and characteristics of cryoproteins

Type I cryoglobulins

- Monoclonal immunoglobulin, often present at high concentrations (>30 g/L)
- Account for ~25% of all cryoproteins
- Occur predominantly with IgM paraproteins (6% of all macroglobulinaemia)
- Especially associated with lymphoma
- Can be seen with monoclonal IgG (particularly of the IgG3 subclass)
- Only occasionally with monoclonal IgA
- Significant cryoprecipitation can occur with low concentrations (~5 g/L)
- Cryoprecipitation occurs most readily in the microvasculature of the colder peripheries (fingers, toes, earlobes, nose); rapid precipitation (24h incubation is usually adequate for their detection)

Type II cryoglobulins

- Monoclonal component (usually IgMκ), usually present at low concentrations (<5 g/L) showing rheumatoid factor activity with polyclonal IgG
- Most often associated with viral hepatitis (hepatitis B and C); can occur with other infections, e.g. mycoplasma pneumonia; can be seen in autoimmune diseases and in lymphoproliferative diseases
- Precipitate slowly (need 72 h incubation)

Type III cryoglobulins

- Polyclonal rheumatoid factor (IgM) and polyclonal IgG or IgA; most often in association with autoimmune rheumatic diseases and persistent infections and only occasionally with lymphoproliferative diseases
- Precipitate slowly (need 72 h incubation)

Cryofibrinogen

- Fibrinogen that precipitates in the cold – collect into EDTA, citrate or oxalate

thermal profile characteristics that influence their clinical significance. Monoclonal proteins that precipitate at temperatures close to 4 °C may not cause any significant problems, but those that precipitate close to room temperature frequently precipitate in the plasma in vivo and are more likely to become clinically manifest, irrespective of relative concentrations.

The clinical features associated with types II and III cryoproteins are linked more to their ability to deposit in tissue as immune complexes rather than to their tendency to precipitate in the cold. Types II and III cryoproteins are both associated with neuropathy, vasculitis and nephropathy.

Immune complex or cryoprotein deposition in tissues is likely to cause activation of the classical complement cascade. Low concentrations of complement (particularly C4) are a good marker of active immune complex disease; measurement of these components is an important part of the investigation of possible cryoproteinaemia. Careful specimen collection and handling are vital. Both serum and plasma samples should be collected, transported and

separated at 37°C. Failure to do this may result in significant cryoproteins being lost in the blood clot during centrifugation.

β_2 -Microglobulin

In conditions in which there is increased cell turnover, for example malignancy (particularly lymphoid), acquired immune deficiency syndromes and inflammatory conditions, plasma β_2 -microglobulin concentrations increase. This protein is cleared from the plasma by glomerular filtration followed by proximal tubular reabsorption and catabolism; plasma concentrations therefore reflect cell turnover and renal function. They have been shown to be an important prognostic indicator in myeloma.

α -Interferon, used in maintenance treatment of some β -Cell malignancies, induces marked increases in plasma β_2 -microglobulin concentrations, and this should be taken into consideration when using β_2 -microglobulin for the assessment of tumour response during α -interferon therapy.

B cell malignancies

Myeloma

Plasma cell disorders have been classified into monoclonal gammopathy of unknown significance (MGUS), smouldering (asymptomatic) multiple myeloma and symptomatic multiple myeloma, based on the presence and concentration of a monoclonal protein, number and morphology of bone marrow plasma cells and end organ damage (hypercalcaemia, renal insufficiency, anaemia, bone lesions attributable to the plasma cell proliferation disorder). Myeloma is the commonest of the malignant causes of paraproteinaemia. The most frequent presenting features are bone pain (in 70% of patients), hypercalcaemia (30%), fever (15%), renal failure (10%) and infection (10%). It is predominantly a disease of the elderly, with a peak incidence between 75 and 80 years.

The three major diagnostic features are the identification of a monoclonal protein in serum or urine, the presence of neoplastic cells in the bone marrow and the destruction of bone. The significance of these and other laboratory investigations is shown in Table 30.31.

As indicated in Table 30.30, IgG, IgA and BJP account for the majority of paraproteins in myeloma. IgD paraproteins occur occasionally, usually with λ light chains. IgM paraproteins are rare in myeloma; IgE paraproteins are extremely rare in any B cell malignancy. The paraprotein concentration at presentation does not predict survival; patients with Bence Jones and IgD paraproteins often present when the paraproteins are present only at relatively low concentrations (they also tend to present at an earlier age) but have shorter survivals than myeloma patients with IgG or IgA paraproteins, even though the concentrations at presentation are often higher in the latter. Bence Jones protein (with or without intact monoclonal immunoglobulin) is seen in over 80% of patients with myeloma; it is rarely found in benign conditions.

The natural history of myeloma is generally one of progression. No treatment can be regarded as curative,

TABLE 30.31 Investigations in myeloma

Investigation	Typical results and their significance
Diagnosis (two of these three are required for a diagnosis of myeloma)	
Serum and urine for presence of paraprotein	Narrow band on electrophoresis shown to be monoclonal by immunofixation
Bone marrow biopsy	Abnormal number and morphology of plasma cells
Skeletal X-ray survey	Lytic lesions (punched out appearance)
Management and prognosis	
Serum calcium	Hypercalcaemia (with normal alkaline phosphatase), due to bone destruction
Serum creatinine and urea	Raised concentrations in renal impairment; ^a the presence of renal impairment is associated with a poor prognosis
Serum albumin	Low concentrations associated with high tumour burden and poor renal function
Full blood count and film	Anaemia, rouleaux formation (due to hyperviscosity), increased background staining (due to paraprotein)
Erythrocyte sedimentation rate (ESR)	High
Non-paraprotein immunoglobulins in serum	Low concentrations due to immune suppression that may predispose to infection
Serum β_2 -microglobulin	High concentrations are associated with high tumour burden and turnover and/or renal impairment and a poor prognosis
Serum paraprotein concentration	In individual patients, correlates with tumour mass so used to monitor response to treatment
Serum free light chains	Particularly useful in non-secretory myeloma and amyloid

^aCauses of renal impairment include deposition of immunoglobulins in the glomeruli, precipitation of light chains causing tubular obstruction, amyloid and hypercalcaemia.

although significant remission may be obtained with newer treatments regimens that may include chemotherapy, steroids, autologous stem cell transplant, bisphosphonates and the newer biologics, lenalidomide, bortezomib and thalidomide. Hypercalcaemia can be a particular problem and cause severe dehydration and/or exacerbate renal failure; it is treated with rehydration and bisphosphonates. Monoclonal protein concentrations that exceed 30 g/L can cause hyperviscosity: this is seen particularly with IgA and IgG3 paraprotein which have a tendency to aggregate. The effects of hyperviscosity can be ameliorated by plasmapheresis. Spinal cord compression requires urgent treatment with dexamethasone, followed by radiotherapy; the latter can also be helpful when localized bone pain is a problem. Infections (often associated with secondary immune deficiency) must be treated promptly. The anaemia may respond to erythropoietin.

Solitary plasmacytoma

In addition to the disseminated form in multiple myeloma, plasma cell tumours may occur as apparently solitary lesions in either bone or soft tissue. Paraprotein concentrations may be very low and intensive urine concentration ($\times 300$) may be necessary to reveal small traces of BJP. Immunofixation of serum, even in the absence of a visible band on electrophoresis, is also worthwhile as the high sensitivity of this technique may reveal a low concentration of paraprotein, which then acts as a valuable marker. Localized treatment by surgery or radiotherapy will frequently result in disappearance of the paraprotein. Patients should be followed up at three-monthly intervals for the first year, six-monthly for the second year and then annually, as there is always a risk that progression to disseminated disease will occur and this may take as long as 20 years from detection of the solitary tumour. Extramedullary deposits may also occur in myeloma, particularly with IgD and Bence Jones type.

Waldenström macroglobulinaemia

IgM myeloma is very rare; IgM paraproteins are more often seen with a heterogeneous group of lymphoproliferative diseases of the B cell lineage such as Waldenström macroglobulinaemia (WM), lymphoma and chronic lymphocytic leukaemia. Waldenström macroglobulinaemia is characterized by the presence of an IgM paraprotein and pleiotropic lymphoid proliferation in the bone marrow. It is generally a disease of elderly men. The presenting symptoms and laboratory investigations in WM are shown in Table 30.32. The malignant cells proliferate

only slowly and are relatively non-invasive; it tends to be the high paraprotein concentration and its effects that account for the presenting features. Hyperviscosity (responsible for the headaches and visual disturbance) is associated with paraprotein concentrations over 30 g/L. Bence Jones protein does occur in WM patients but tends to be at lower concentrations and less damaging than in myeloma.

Lymphomas, chronic lymphocytic leukaemia and heavy chain diseases

A paraprotein (most often IgM) is found in 10–20% of patients with non-Hodgkin lymphoma (NHL) and 5–15% of patients with chronic lymphocytic leukaemia (CLL). These tumours arise from cells early in the B cell maturation and the monoclonal protein concentration is usually low. The tumours are diagnosed by biopsy and histological examination of lymphoid tissue and by immunophenotyping of blood and lymph node biopsy material.

Certain B-lymphoproliferative disorders are associated with synthesis of heavy chains or heavy chain fragments in the absence of light chain synthesis; most often the fragments consist of heavy chains with deletion of the variable and first constant region domain. Free heavy chains are more susceptible to enzymatic degradation than intact immunoglobulin; they undergo considerable post-synthetic modification and migrate in electrophoretic separations as rather diffuse zones, anywhere between the α_1 and fast γ zone. Low concentrations of the heavy chain paraprotein may be missed on examination of the electrophoretic separation. The presence of heavy chains can be indicated by immunofixation when reaction with heavy chain, but not light chain, antiserum is seen. Identity is confirmed by an immunoselection technique whereby the serum is electrophoresed through an agarose gel layer containing antiserum to light chains where any complete immunoglobulin will be trapped; free heavy chains pass through the light chain 'trapping' layer then move into a second layer containing antiserum to the appropriate heavy chain (α , γ or μ) and precipitate as they travel through this layer to form a 'rocket'. The rocket is usually directly visible but can be stained with protein stains if necessary. It is essential that a positive heavy chain sample and a myeloma protein of the same class are included as controls in the heavy chain immunoselection technique. Bence Jones protein is unusual in α - and γ -chain diseases and is more often seen with μ -chain disease. Immunosuppression of non-paraprotein immunoglobulin is frequent in the heavy chain diseases.

α -Chain disease represents the most distinct clinical entity among the heavy chain diseases. The disease is a lymphoma, usually of the gut, and is alternatively known as Mediterranean type lymphoma. It is interesting in that the initial diffuse infiltration appears benign and complete remission can be achieved in some patients with oral antibiotics. However, untreated, the condition progresses to frank malignancy. The diagnosis depends on the finding of α -chains in the serum. Gut biopsies show villus

TABLE 30.32 Clinical features and results of investigations in Waldenström macroglobulinaemia

Clinical feature	% of patients
Weakness and fatigue	41
Hepatosplenomegaly	40
Lymphadenopathy	36
Bleeding tendency	31
Weight loss	18
Neurological	12
Visual disturbances	10
Infection	8
Bone pain and arthralgia	4
Raynaud phenomenon	3
Investigation	Typical results
Serum and urine for presence of paraprotein	Narrow band on electrophoresis shown to be monoclonal on immunofixation
Bone marrow biopsy	Increased number of lymphocytoid cells
Plasma viscosity	Increased
Lymph node biopsy	Well preserved architecture with lymphocytoid infiltrate
Haematology investigations	Anaemia, clotting disorders, rouleaux formation, high ESR

atrophy and plasma cell infiltration of the lamina propria. Clinical features include a malabsorption syndrome and diarrhoea.

γ -Chain disease is not characterized as a distinct clinical entity but has a varied clinical and pathological picture ranging from apparently non-malignant proliferation, for example systemic lupus erythematosus, to aggressive lymphoma. An association with autoimmune diseases has been recorded in about 25% of published cases, most frequently with rheumatoid arthritis and autoimmune haemolytic anaemia and, in a few cases, malignant NHL that was not diagnosed until many years after the autoimmune disorder.

μ -Chain disease is extremely rare and most often a variant of CLL.

Monoclonal gammopathy of unknown significance (MGUS)

A clearer definition of MGUS is gradually emerging as a serum monoclonal protein present at a concentration <30 g/L, clonal bone marrow plasma cells $<10\%$ and no end-organ damage (e.g. hypercalcaemia, renal insufficiency, anaemia or bone lesions). Nevertheless, the finding of a paraprotein without any symptoms poses a problem, and a relatively frequent one, given that sensitive electrophoretic techniques are so readily available and there are few constraints on requesting such investigations. Benign disease can be indistinguishable from malignant disease early in its course. No single feature can help differentiate these two entities, but monoclonal protein concentration, the presence and degree of any immune suppression and the presence of BJP can give some guide. Thus, in general, benign disease is seen with serum paraprotein concentrations below <10 g/L (not applicable for IgD or BJP only): higher concentrations, and particularly values >30 g/L, suggest malignancy. Non-paraprotein immunoglobulins are suppressed in 75% of myeloma patients at presentation. The presence or absence of BJP is the most important distinguishing feature; myeloma should be considered in any patient with BJP in the urine at a concentration of >10 mg/L. The International Myeloma Working Group suggests that the measurement of serum free light chains can help to stratify patients with MGUS. The risk of progression of MGUS to an overt malignancy has a cumulative probability of 10% at 10 years (approximately 1% per year). The follow-up time in patients with MGUS should be determined by the clinical findings and the patient's age. A general guide would be to examine the paraprotein in serum and urine at three-monthly intervals for the first year, six-monthly for the second year and, if no progression occurs, annually thereafter. If any other features of B cell malignancy occur, full investigation is required.

Transient paraproteinaemia

Transient paraproteins are usually small IgG or IgM bands (~ 1 – 2 g/L); however, they may be >10 g/L. They typically occur during infections, often in patients with compromised immune systems, for example patients with CLL or with iatrogenic immune suppression for

organ transplantation. Transient paraproteins may persist for over one year but most disappear within a few months.

Serum free light chains (SFLC)

Methods for the quantification of serum free light chains are now widely available, with more manufacturers producing reagents. The assays use antibodies directed against the 'hidden' epitopes of free light chain molecules at the interface between the light and heavy chains of intact immunoglobulins. The general concerns about the immunochemical quantification of a monoclonal protein are as applicable to SFLC as they are to the quantification of intact monoclonal IgG, IgA or IgM. However, there are additional concerns about the FLC assays with respect to their linearity, performance in external quality assurance schemes, reproducibility between laboratories and quantification of the proteins. Nevertheless, there is clear value to these measurements in patients with light chain amyloidosis and in non-secretory myeloma where the serum monoclonal protein may be undetectable by other methods. They are being used as 'stringent' criteria for remission in some trials investigating new treatment regimes in myeloma. There are also publications promoting their use to stratify risk in myeloma, other B cell dyscrasias and MGUS.

Amyloidosis

Amyloidosis is a heterogeneous group of clinical conditions characterized by the systemic or localized deposition of fibrils derived from a variety of protein precursors and having characteristic histological appearance. It can be systemic or localized, and both types can be acquired or hereditary (Table 30.33). This table also indicates the nature of the deposit and precursor proteins, where known. In addition to the precursor proteins, amyloid deposits can contain sulphated glycosaminoglycans and serum amyloid P component, which is a normal circulating protein in human plasma.

The tendency for monoclonal immunoglobulin to form amyloid fibrils (AL) can occur as a secondary complication of B cell malignancies. Amyloidosis occurs in 6–15% of patients with myeloma and is more common when the major paraprotein is free light chains. When amyloid fibrils are found in the absence of myeloma (previously classified as primary idiopathic amyloidosis), 90% of patients have a demonstrable serum paraprotein or Bence Jones proteinuria, albeit often at low concentrations; in addition, there is frequently a degree of hypogammaglobulinaemia and a slight increase in plasma cells in the bone marrow. Progressive amyloid deposition is the most serious clinical complication in these patients as they do not tend to progress to overt malignancy.

The clinical features of peripheral neuropathy, cardiomyopathy, purpura, macroglossia and carpal tunnel syndrome are similar in both the malignant and benign conditions and both have similar age distribution (mean 55–62 years), with males more frequently affected than females.

TABLE 30.33 Amyloidosis

	Disease/association	Fibril	Precursor	
Systemic	Acquired	Paraproteinaemia	Monoclonal Ig light chain	
		Chronic inflammation/infection	SAA	
		Haemodialysis	Raised plasma β_2 M	
		Senile type	Plasma PA	
	Hereditary	Neuropathic		
		Type I (Portuguese, Japanese, Swedish, Jewish)	PA	Plasma PA
		Type II (Swiss, German, Indiana-Maryland)	PA	Plasma PA
		Type III + nephropathy (Irish, Scottish, English, Iowa)	Apolipoprotein A1	
		Type IV (Finnish)	?	
		Nephropathic		
Familial Mediterranean fever	AA	SAA		
Others (Ostertag, Muckle–Wells)	?			
Cardiomyopathic	PA	Plasma PA		
Localized	Acquired	Cerebral (Alzheimer, senile dementia, Down syndrome)	Precursor encoded on chromosome 21	
		Endocrine (APUDomas)	Calcitonin	
		Nodular deposits		
		Primary cutaneous		
	Hereditary	Ocular deposits	AL	Monoclonal Ig light chains
		Cerebral	?	? Keratin
		Icelandic type	?	
		Dutch type	Cystatin C	Cystatin C
			B-protein	Precursor encoded on chromosome 21
		Cutaneous	?	

APUD, amine precursor uptake and decarboxylation; ASc₁, an amyloid protein derived from prealbumin; β_2 M, β_2 -microglobulin; PA, prealbumin; SAA, serum amyloid A protein.

While the detection of a serum or urine paraprotein is an important finding in AL amyloidosis, the diagnosis depends on demonstration of amyloid fibrils in biopsy material. Tissue, usually rectal mucosa or abdominal fat, is stained with Congo red: the presence of characteristic apple green birefringence under polarized light is diagnostic. The National Amyloidosis Centre can localize amyloid deposits using gamma camera photographs after injection of radiolabelled serum amyloid P material. Increasingly, serum free light chain measurements are used to monitor patients but other biochemical markers (e.g. the cardiac biomarkers troponin-T (cTnT) and N-terminal pro-B-type natriuretic peptide (NT-proBNP) in cardiac amyloid) are equally important.

INFECTION AND SEPSIS

All bacteria are recognized as foreign and cause inflammation on entry into tissues. This is achieved by various mechanisms, many of which are innate. In addition, many components of bacteria are able to stimulate inflammatory or immune responses, and some of these are shown in Table 30.34.

Infection with a small bacterial load, particularly if confined to a local site, will cause cytokine production, a controlled inflammatory response and recruitment of immune cells. The outcome will be the resolution of the infection, the repair of any damage to surrounding tissue and wound healing. If the bacterial load is

TABLE 30.34 Proinflammatory bacterial components and their modes of action

Component	Mode of action
Carbohydrates in cell walls (O-antigen of lipopolysaccharide (LPS))	Activation of complement via alternative pathway
Endotoxin (Gram-ve organisms)	Engage macrophage receptors (CD14) to induce inflammatory cytokine production (mainly IL-1, IL-6 and TNF α)
Peptidoglycan (Gram +ve organisms)	Activate T cells, production of immune cytokines
Superantigens, e.g. enterotoxin of <i>Staphylococcus aureus</i>	
Exotoxins, e.g. anthrax lethal toxin	Enzymic properties induce cytokines via various mechanisms

greater, the bacteria and/or the cytokine response may not be confined and may extend to produce a generalized systemic response. If this is unabated, it will eventually lead to septic shock, the condition in which vital tissues are inadequately perfused with blood, with progressive deterioration of circulation and falling blood pressure leading inevitably to death if untreated. Because the inflammation progresses through cascades of host-mediated reactions, once the inflammatory

response has been initiated, stopping it becomes increasingly difficult – even removal of the bacteria may not affect the outcome.

Diagnosis and monitoring of infections

The diagnosis of infection is based on clinical findings, but microbiological investigation is often required to identify the organism responsible and information about antibiotic resistance may help guide treatment. However, there are situations when the clinical signs of infection may be masked, for example in patients who are immunosuppressed. In these situations, markers of acute phase response and, in particular, measurement of C-reactive protein (CRP) concentrations may be useful to give an independent marker of the presence of infection or inflammation and to monitor the response to treatment.

C-reactive protein and markers of the acute phase response

Measurement of the erythrocyte sedimentation rate (ESR) is one of the oldest ways of detecting the acute phase response. It is still frequently requested but has important limitations, including its slow response and insensitivity to small changes in disease activity. Plasma viscosity measurements have also been used to monitor the acute phase response but suffer from the same limitations as the ESR.

C-reactive protein is produced by the liver on stimulation with the inflammatory cytokines IL-6, IL-1 and TNF α . There is a rapid increase in CRP concentrations, with changes detectable as quickly as 6 h after the inflammatory insult, with a half-life of 19 h. C-reactive protein is the most widely measured acute phase protein, with raised concentrations occurring in both bacterial infections and inflammation. A raised CRP concentration is unequivocal evidence of an inflammatory response. However, viral infections do not usually cause a raised CRP and neither do some autoimmune diseases, for example SLE and scleroderma. Minimal inflammation causes a significant increase in CRP concentration but most analytical methods lack the sensitivity to detect this. There are also marked variations in quoted reference ranges, for example, <0.1 mg/L, <1 mg/L and <10 mg/L. The concentration of CRP is related to the extent and severity of inflammation, with concentrations of 10–40 mg/L occurring in mild inflammation, 40–200 mg/L in acute inflammation and bacterial infections, and >300 mg/L in extensive trauma and severe sepsis. In patients with bacterial infections, daily measurement of CRP may be useful to monitor response to treatment. Appropriate antibiotic therapy should result in the CRP falling by approximately 50% every 24 h. If the fall in CRP is slower than this, it suggests that the antibiotic regimen may not be appropriate. Patients with severe systemic inflammatory responses, for example the critically ill, often have infection on top of a marked inflammatory response, and in these situations, a fall in CRP of >20% in 24 h is consistent with successful treatment of the infection,

TABLE 30.35 Clinical applications of C-reactive protein measurements

Disease category	
? Unwell	Raised CRP concentration is unequivocal evidence of inflammation. Normal CRP concentration does not exclude significant inflammation
Bacterial infection	Major elevations normally seen in bacterial infections Useful to monitor response to antibiotic treatment
Viral infections	Only minor elevations seen – unless complicated by secondary bacterial infection
Fungal infections	Only minor elevations with localized infections; marked elevations when systemic
Transplantation	Raised concentrations reported during organ rejection episodes – but also in infection Raised concentrations help to distinguish bacterial infections from graft-versus-host disease (GVHD) in bone marrow
Chronic inflammatory states	Raised concentrations in many inflammatory conditions, therefore diagnostically useful Useful to monitor disease progression and treatment (RA particularly) Useful to detect intercurrent infection (SLE particularly)

whereas a rise in CRP of >20% in 24 h is consistent with a new infection. In chronic inflammatory diseases, for example rheumatoid arthritis, a falling CRP concentration precedes clinical and radiological improvements in disease.

Some laboratories are measuring CRP using assays with improved sensitivity (0.1–5 mg/L, so-called ‘highly-sensitive CRP’) for use as a marker of chronic low-grade inflammation to predict risk of coronary heart disease and progression of unstable angina. The combination of high analytical imprecision and high biological variation at low concentrations makes this difficult in a ‘one-off’ analysis, but if the analytical systems improve, CRP may become more widely applicable in predicting risk of cardiovascular disease. The major clinical applications of CRP measurement are shown in [Table 30.35](#).

Procalcitonin (PCT) is a peptide that undergoes post-translational proteolysis into the mature hormone calcitonin. High plasma PCT concentrations are seen in patients with severe bacterial and fungal infections but also during the acute phase response post surgery. There is considerable interest in the possible use of this protein as a marker of acute sepsis, particularly in neonates, but it has not been shown to be clearly superior to CRP.

TRANSPLANTATION

Transplantation of organs or tissues is a complex procedure undertaken to restore a damaged or missing function, or to replace a lost, removed or absent 'tissue'. Transplants may be autologous (within same person, e.g. stem cell transplant post-chemotherapy), syngeneic (identical genetics, i.e. from a twin) or allogeneic (from a different person). Table 30.36 shows the major types of transplantation currently performed in the UK, and some of the requirements applicable to each type of transplant. (General information about transplantation can be found at: www.histocompatibilityandimmunogenetics.com)

Immunological issues in transplantation include rejection due to the graft being recognized by the immune system as foreign, graft-versus-host disease (GvHD) and infection secondary to the immune suppression required to prevent or control rejection or GvHD (see Box 30.4). The key issue, however, remains rejection of the graft. This can happen by three different mechanisms, shown in Box 30.5. The risk of rejection or GvHD can be reduced by matching the donor and recipient for ABO blood group and HLA type. This is particularly important for renal and stem cell transplants. A longer-term complication of modulating the immune system to accommodate transplanted tissue is an increased risk of malignancy.

Organ transplantation

Patients waiting for an organ transplant are registered, together with their HLA type, on a waiting list. As potential donors become available their HLA type is compared with those of the patients on the waiting list. Potential donor-recipient matches need to be further investigated using a specific test of the reactivity of the recipient's serum against the donor. This is particularly to check whether the recipient has any pre-formed antibodies to the donor that could cause hyperacute rejection. Transplant patients usually need long-term immunosuppression to prevent graft rejection; the most common immunosuppressants with their modes of action are shown in Table 30.37.

TABLE 30.36 Organ transplantation in the UK, April 2011–March 2012

Organ/tissue	Number in UK
Kidney	2801
Heart/lung	321
Liver	791
Pancreas	249
Cornea	3520
Intestine	22
Stem cell transplantation (prepared from peripheral blood, umbilical cord or bone marrow)	Not available

Taken from NHS Blood and Transplant Activity Report 2011–2012: www.organdonation.nhs.uk

BOX 30.4 Immunological complications of transplantation

1. **Rejection:** controlled by HLA and blood group matching and immunosuppression. For 'solid' organ transplantation, the aim is to prevent the recipient immune system from attacking the transplanted organ. With stem cell transplantation, the recipient's immune system will have been destroyed by conditioning, thus reducing the risk of rejection.
2. **Graft versus Host Disease (GvHD):** the main cause of morbidity and mortality post-stem cell transplantation. Donor immune cells attack the recipient, particularly in the skin, gut and liver. The severity depends on the degree of HLA matching but the risk is reduced by T cell depletion of the stem cell preparation before infusion into the donor. Some T cells may have a graft versus leukaemia (GvL) effect, which can be beneficial.
3. **Infection secondary to immune suppression:** Suppressing the immune system to prevent rejection or GvHD increases the risk of infection. Suspected infections must be treated promptly with broad spectrum antibiotics, antiviral or antifungal drugs. In stem cell transplantation, conditioning causes profound myelosuppression, so cell counts, particularly of neutrophils and platelets, are monitored closely as immune reconstitution by the donor stem cells takes place. Additional infection control measures are used, including nursing the patient in air-filtered isolation room and providing specially sterilized food until the neutrophil count is $>0.5 \times 10^9$ cells/L. Prophylactic antibiotics, antivirals and anti-fungals may also be used to minimize the risk of infection with organisms such cytomegalovirus (CMV), Epstein-Barr virus (EBV), adenovirus, *Candida* and *Aspergillus*.

BOX 30.5 Types of rejection

- **Hyperacute rejection:** this is caused by the recipient having pre-formed antibodies to the donor's cells. This can be due to previous blood transfusion, multiple pregnancies or transplantation and the antibodies can be to blood group or HLA antigens. The antibodies bind to their antigenic target in the graft blood vessels as the graft is perfused by the recipient's blood. This activates the complement cascade, damages the blood vessels, activates platelets and causes blockage of the graft microvasculature, resulting in graft failure in minutes to hours. The risk can be minimized by HLA typing and pre-transplant cross-matching.
- **Acute rejection:** this is caused by recipient T cells recognizing donor antigen presented by antigen-presenting cells in the draining lymph nodes of the graft. The T cells attack the graft, resulting in graft failure within days-weeks. Production of anti-donor antibodies may also be a consequence of this process compounding the damage by an antibody-mediated response. Compatible HLA typing minimizes the risk of acute rejection but the immune suppression is focused on preventing it.
- **Chronic rejection:** the cause is poorly understood but there is gradual (over months to years) occlusion of blood vessels resulting in local tissue necrosis and ultimately graft rejection. Cellular and antibody mechanisms are both thought to contribute.

TABLE 30.37 Major immunosuppressive drugs and their actions

Drug	Action	Comments
Ciclosporin A	Inhibits T cell activation	Nephrotoxic, so minimize use in renal transplant Serum concentrations monitored
Tacrolimus (FK506)	Inhibits T cell activation	Nephrotoxic, so minimize use in renal transplant Serum concentrations monitored
Sirolimus (rapamycin)	Inhibits cell proliferation by blocking cell cycle	
Steroids	Suppress expression of inflammatory genes, e.g. cytokines	
Azathioprine	Purine analogue that inhibits proliferation	
Mycophenolate mofetil (MMF)	Inhibits proliferation by targeting inosine monophosphate dehydrogenase	

TABLE 30.38 Sources of stem cells

Source of stem cells	Cord blood	Peripheral blood	Bone marrow
Preparation	None	GCSF is used to mobilize stem cells from bone marrow into blood – small risk of malignancy	None
Collection	Painless; would be wasted otherwise	Minimal discomfort; no hospital stay or anaesthesia	Painful collection Requires hospital admission General anaesthesia
Storage	Can be stored, so readily available when required	Collected on day of transplant, so requires coordination	Collected on day of transplant, so requires coordination
Yield (CD34+ cells/kg)	Approx. $0.5-1 \times 10^6$	Approx. 7×10^6	Approx. $2-3 \times 10^6$
Function	Does not confer immunity as immunologically naive – reduced risk of GvHD but lack of GvL	Can confer immunity	Can confer immunity

GCSF, granulocyte colony stimulating factor; GvHD, graft-versus-host disease; GvL, graft-versus-leukaemia.

Stem cell transplantation

This is the re-population of the bone marrow to rescue and restore an immune system ablated as a result of high dose chemo- or radiotherapy or to treat genetic diseases. Stem cells are undifferentiated cells that are capable of self-renewal and differentiation into more specialized cells. Stem cells can be harvested for transplant from cord blood or bone marrow but are now most commonly obtained from peripheral blood (see Table 30.38). The relative ease with which stem cells can be collected from peripheral blood and safely stored has made autologous (from the patient) peripheral blood stem cell transplant (PBSCT) a common 'rescue' treatment following high dose chemotherapy, e.g. for malignancy. Allogeneic (from a donor) stem cell transplants may be required to treat haematological malignancies, e.g. leukaemias, and bone marrow disorders such as thalassaemias, aplastic anaemia and severe combined immunodeficiency.

There are registers for people willing to be bone marrow or stem cell donors, which contain details of their HLA types. When a patient requires an allogeneic stem cell transplant (and a suitable family member is not available), their HLA type is used to search donor

databases to try to find an appropriate match. If a donor is identified, their HLA type is double-checked and the potential recipient's sample cross-matched with the potential donor. The transplant typically involves 'conditioning' with chemotherapy or radiotherapy to eradicate disease (depending on the reason for transplant) and the recipient's functioning bone marrow. The donor stem cells are infused like a blood transfusion; they 'home in' to the bone marrow where they start to develop and multiply. The patients are immunosuppressed to prevent rejection and GvHD. The main immunological complications of PBSCT are shown in Box 30.5.

CONCLUSION

The immune system is complex and the diseases associated with abnormalities of the immune system can affect almost every part of the body. The common symptoms where the immune system is involved are shown in Table 30.39, together with the appropriate investigations and relevant disease. The list is not exhaustive but covers the most important immunological investigations.

TABLE 30.39 Clinical features and investigations in immunological diseases

Organ or clinical feature	Investigation(s)	Disease, pathological process or comments
Recurrent infections	Immunoglobulins (IgG, IgA and IgM)	Primary immune deficiency, e.g. common variable immunodeficiency (CVID) Secondary immune deficiency, e.g. B cell malignancy (+ urine for BJP)
	CRP	To support diagnosis and monitor response to treatment
	Complement (CH50)	Complement deficiency – Neisserial infections, meningitis
Liver	α_1 -Antitrypsin	α_1 -Antitrypsin deficiency (especially in neonates)
	Antinuclear antibodies	Autoimmune hepatitis
	Smooth muscle antibodies Mitochondrial antibodies	Autoimmune hepatitis Primary biliary cirrhosis (M2 variant, IgM concentrations usually raised)
Kidney	Antinuclear antibodies	Systemic lupus erythematosus (SLE)
	Antinuclear cytoplasmic antibodies	Vasculitis
	Glomerular basement membrane antibodies	Goodpasture syndrome (GBM disease)
	Serum immunoglobulins and serum and urine electrophoresis	B cell malignancy (with renal damage due to AL amyloid, cast formation etc.)
	Cryoglobulins	Immune complex deposition in the kidneys causing renal impairment
Respiratory system	C3 and C4	Immune complex disease and nephritis (SLE, cryoproteins etc.)
	Total and specific IgE	Allergy
Anaphylaxis	Antinuclear cytoplasmic antibodies	Vasculitis
	GBM antibodies	Goodpasture syndrome (GBM disease)
	α_1 -Antitrypsin	COPD
Skin	Serum tryptase	Type I hypersensitivity
	Total and specific IgE	Allergy
Skin	C1 esterase inhibitor, C3 and C4	Hereditary angioedema
	Antibodies to skin basement membrane and intercellular cement	Pemphigus and pemphigoid
	Cryoglobulins (with serum and urine paraprotein studies)	Deposition of cryoproteins in the skin because of their precipitation in the cold Deposition of immune complexes in the skin causing vasculitis
	C1 esterase inhibitor, C3 and C4	Hereditary angioedema
	Antinuclear antibody	Urticaria SLE Systemic sclerosis
Gut	Total and specific IgE	Allergy
	Total and specific IgE	Food allergy
	C1 esterase inhibitor, C3 and C4	Hereditary angioedema
	Immunoglobulins (IgG, IgA and IgM)	Immune deficiency
	Endomysial antibodies (or anti-tissue transglutaminase antibodies)	Coeliac disease
Bone and joint pain	Serum and urine paraprotein studies	Myeloma
	Rheumatoid factor	Rheumatoid arthritis (CRP for monitoring)
	Anti-cyclic citrullinated peptide (CCP) antibodies	Earlier predictor of rheumatoid arthritis
Thyroid	Antinuclear antibodies	SLE
	Anti-thyroid peroxidase antibodies	Present in autoimmune hypothyroidism
Pancreas	TSH receptor antibodies	Present in Graves disease
	Glutamic acid decarboxylase antibodies	Present in type I diabetes
Anaemia	Islet cell antibodies	Present in type I diabetes
	Serum and urine paraprotein studies	Myeloma
	Endomysial antibodies (or anti-tissue transglutaminase antibodies)	Coeliac disease
	Gastric parietal cell antibodies and intrinsic factor antibodies	Pernicious anaemia

Further reading

The following texts provide good accounts of more specialized areas of immunology.

- Dimopoulos M, Kyle R, Fermand J-P et al. Consensus recommendations for standard investigative workup: report of the International Myeloma Workshop Consensus Panel 3. *Blood* 2011;117:4701–5.
- Gahrton G, Durie BGM, Samsom DM, editors. Multiple myeloma and related disorders. London: Arnold; 2004.

- Gompels MM, Lock RJ, Abinum M et al. C1 inhibitor deficiency: consensus document. *Clin Exp Immunol* 2005;139:379–94.
- Graziani M, Merlini G, Pettrini C. Guidelines for the analysis of Bence Jones protein. *Clin Chem Lab Med* 2003;41:338–46.
- International Union of Immunological Societies Expert Committee on Primary Immunodeficiencies. Primary immunodeficiencies: 2009 update. *J Allergy Clin Immunol* 2009;124:1161–78.
- Keren DF. Protein electrophoresis in clinical diagnosis. London: Arnold; 2003.

- Milford Ward A, Sheldon J, Rowbottom A et al. editors. PRU handbook of clinical immunochemistry. 9th ed. Sheffield: PRU Publishing; 2007.
- Milford Ward A, Sheldon J, Wild DG, editors. PRU handbook of autoimmunity. 10th ed. Sheffield: PRU Publishing; 2007.
- Morgan BP. Complement. In: Collier L, Balows A, Sussman M, editors. 10th ed. Topley and Wilson's microbiology and microbial infections, vol. 3. London: Hodder Arnold; 2004. p. 1–23.
- Mygind N, Dahl R, Pedersen S et al. editors. Essential allergy. 2nd ed. Oxford: Blackwell Scientific; 1996.
- Peakman M, Vergani D. Basic and clinical immunology. 2nd ed. Edinburgh: Churchill Livingstone; 2008.
- An excellent textbook of immunology that will give a good foundation to the subject.*
- Rich R, Fleisher TA, Shearer WT et al. editors. Clinical immunology: principles and practice. In: 3rd ed. London: Mosby; 2008.
- A comprehensive account of the subject at postgraduate level.*
- Rose NR, Mackay IR. The autoimmune diseases. London: Academic Press; 1998.
- Spickett G. Oxford handbook of clinical immunology. 3rd ed. Oxford: Oxford University Press; 2013.
- An up-to-date 'ready reference' summary.*

APPENDIX 30.1: IMMUNOLOGICAL INVESTIGATIONS

Quantitative assays should be calibrated against available international reference preparations; be sensitive, specific, precise and where available, validated by adequate performance in external quality assurance (QA) schemes. It is important to mention these factors because there is a trend towards moving immunoglobulins and complement measurements from dedicated immunoassay analysers to main clinical chemistry analysers and we must not compromise our ability to detect immune deficiencies or the overall quality of the assays by losing precision and sensitivity. It is also worth remembering that babies and children will be a group of patients often investigated for immune deficiency and saving any 'left-over' samples for further analysis can be very useful.

Quantification of total immunoglobulin concentrations

Methods for the quantification of serum IgG, IgA and IgM concentrations are readily available. Quantification of serum immunoglobulin concentrations should be accompanied by serum protein electrophoresis and, in adults, by urine protein electrophoresis.

There is marked variation in immunoglobulin concentrations with age, and results must be interpreted using appropriate age-related reference ranges and the potential impact of transfusion of any blood products, maternal IgG and prematurity.

IgG subclasses

The IgG1 and IgG3 subclass concentrations mature first, with IgG2 and IgG4 concentrations maturing more slowly. There is considerable debate among immunologists about the value of IgG subclass measurements. It is likely that many IgG subclass requests are of negligible value; e.g. a patient with a low total serum IgG will have low IgG concentrations of subclasses. There may be a small number of

patients where the subclass measurements may be useful but increasingly, quantitative specific antibody responses are being used.

Quantification of specific antibody responses

These assays quantify the antigen-specific IgG. Ideally, every individual should be vaccinated against tetanus and should have measurable IgG concentrations to tetanus. There are assays for IgG to *Haemophilus influenzae* B, *Pneumococcus* and tetanus but the assays can be variable and the external QA programmes show marked variability in performance. Analysis of pre- and post-vaccination samples helps to limit the effect of inter-assay variability. Patients with a low specific antibody concentration should be re-vaccinated and the concentrations re-checked 4–6 weeks later. Unfortunately, antibody production and concentration are not necessarily good correlates of immunological protection so the interpretation of the results needs particular care.

Quantification of IgE

Reliable assays for IgE are becoming increasingly available on immunoassay analysers. It is interesting to note that in the investigation of immune deficiency (e.g. hyper IgE syndrome), immunologists now tend to look for raised IgE concentrations (into many thousands of kU/L) rather than low concentrations.

Complement

It is possible to quantify all the complement components but only C3 and C4 assays are readily available. Concentrations of other proteins, e.g. C1 esterase inhibitor can be easily measured but tend only to be available in more specialized centres. The functional assays are also vital in the investigation of immune deficiencies; these are variations on total haemolytic complement assays (for the classical and alternative pathways).

Enumeration of cell numbers

The total white cell count and differential is a vital investigation that will include total lymphocyte count and total neutrophil count but it cannot, for example, distinguish T lymphocytes from B lymphocytes. The various white cell types can be defined using flow cytometric assays that look at the size and granularity of the cells and the binding of fluorescent conjugated antibodies to the cell surface markers. The assays are usually done in immunology or haematology departments. Table 30.3 shows the common CD antigens, many of which are used in the investigation of suspected immune deficiency. Reference ranges are age-related and the results reported as either percentages or as absolute numbers.

Functional assays

There are immune deficiencies, where the various cell types are present but they lack any functional ability. Assays of lymphocyte function are laborious and complex and

typically rely on the proliferation of cells after stimulation with mitogens, antigens or stimulation via specific receptors. These assays are often run over 3–5 days, so need careful planning, and the interpretation of the results needs great expertise, therefore they are only done in specialized centres.

Neutrophil function tests

Neutrophil function tests are also complex and like lymphocyte function tests, the cells need to be alive for them to work. The most common analysis is for detection of the neutrophil respiratory burst, either by a flow cytometric assay (using dihydrorhodamine) or by looking at

neutrophils under the microscope after being incubated with the redox dye, nitroblue tetrazolium.

Autoantibodies

The analytical problems with autoantibody testing are many and include the source, type and preparation of antigen, the method used and variation between each patient's antibodies. Autoimmune serology, in comparison with general clinical biochemistry, is poorly standardized and this is compounded by large numbers of companies producing kit reagents for these tests. External quality assurance schemes are in place but they can make scary reading for a laboratory scientist!

Metabolic bone disease

Timothy Cundy • Ian R. Reid • Andrew Grey

CHAPTER OUTLINE

BONE BIOLOGY 604

- Anatomy of bone 604
- Bone matrix proteins 605
- Cellular elements of bone 607
- Biochemical markers of bone turnover 609

OSTEOPOROSIS 613

- Causes of osteoporosis 614
- Investigation and diagnosis 615
- Treatment 617

OSTEOMALACIA 620

- Calciopenic osteomalacia 620
- Phosphopenic osteomalacia 622
- Osteomalacia and acidosis 623
- Defective osteoblast function and osteomalacia 623

CHRONIC KIDNEY DISEASE – MINERAL AND BONE DISORDER 624

- Aetiology 624
- Clinical features 625
- Investigations 625
- Treatment 627
- Bone disease after renal transplantation 628

BONE DISEASE IN PRIMARY HYPERPARATHYROIDISM 628

- Clinical, biochemical and histological features 628
- Treatment 629

PAGET DISEASE OF BONE 629

- Epidemiology 629
- Aetiology 629
- Natural history 629
- Pathology 630
- Clinical features 630
- Investigations 630
- Responses to treatment 631

BONE TURNOVER AND BONE DISEASE IN CHILDREN 632

- GENETIC BONE DISEASES 632
- Osteogenesis imperfecta 632
- High bone mass 634
- Other disorders 634

CONCLUSION 634

APPENDICES 635

BONE BIOLOGY

The principal role of the skeleton is a structural one, maintaining body shape, providing protection for internal organs and, together with the neuromuscular system, making locomotion possible. It also has an important secondary role in mineral homeostasis, functioning as a reservoir for calcium ions in particular. Metabolic bone diseases can affect both these functions.

Anatomy of bone

Macroscopic

The anatomist classifies bones as being either flat (e.g. skull, scapula, mandible, ilium) or long (e.g. the limb bones). Flat bones result from intramembranous ossification; long bones predominantly from endochondral ossification. A long bone consists of a shaft (diaphysis) broadening at either end into an epiphysis. The transitional zone between the diaphysis and the epiphysis is termed the metaphysis. On sectioning a long bone, two patterns of organization of bone tissue are found. The

elements of bone can be packed together without intervening marrow spaces to form cortical or compact bone, or they can form an interlacing meshwork of trabeculae referred to as cancellous or trabecular bone. The diaphysis of the long bone consists mainly of cortical bone, whereas the metaphysis and epiphysis have a greater quantity of trabecular bone, enclosed within a thin cortical envelope. Some 80% of the weight of an adult human skeleton consists of cortical bone. However, the surface-to-volume ratio of trabecular bone is very much higher than that of cortical bone and it is metabolically much more active.

Microscopic

At a microscopic level, bone consists of matrix (~35% by volume), mineral (~60%) and cells (<5%). The matrix is predominantly type I collagen fibres, usually organized in layers within which the fibres are parallel to one another. In adult bone, the fibre orientation varies from one layer to the next and this is referred to as lamellar bone. If deposited along a flat surface, the lamellae will be parallel to that surface, but in cortical bone they are

concentrically oriented around a central blood vessel to form the Haversian canal system. When bone formation is rapid (e.g. during growth or fracture healing) collagen fibres may be laid down with more random orientation, producing woven bone.

The mineral phase of bone is hydroxyapatite ($\text{Ca}_{10}(\text{PO}_4)_6(\text{OH})_2$). This forms spindle-shaped crystals, which are found in association with the collagen and ground substance. Their orientation is usually parallel to that of the collagen fibres.

There are two general cell types in bone: osteoblasts and osteoclasts. Both are found on the bone surface at sites of active remodelling. Osteoblasts are also thought to be the precursor of two cell types that are more widespread: bone lining cells, which are found over inactive bone surfaces, and osteocytes, which are found in lacunae scattered throughout bone. They are thought to be osteoblasts that have been engulfed by the bone that they have formed. They have long cell processes, which are in contact with similar processes arising from other osteocytes or with those of the bone lining cells. The cell processes lie in an interconnecting network of canaliculi that extends throughout bone tissue. The bone lining cells and osteocytes thus delineate an extracellular fluid (ECF) space that is in contact with the bone surface. This space has a volume of 1–1.5 L and a surface area of several thousand square metres. This is the site of mineral exchange between ECF and bone. The size and nature of this exchange is unknown but the fluid within this space, the bone ECF, has an ionized calcium concentration of only 0.5 mmol/L – less than half of that elsewhere in the ECF.

Bone matrix proteins

Collagen

Almost 90% of the protein in bone matrix is type I collagen, which is synthesized by osteoblasts. It is a large molecule (MW >300 000 Da) with a trimeric helical structure. Type I collagen is initially synthesized in the rough endoplasmic reticulum (RER) as a precursor molecule (type I procollagen) that combines two $\text{pro}\alpha 1(\text{I})$ and one $\text{pro}\alpha 2(\text{I})$ peptide chains (coded by *COL1A1* and *COL1A2*, respectively) in a triple helix. $\text{Pro}\alpha 1(\text{I})$ and $\text{pro}\alpha 2(\text{I})$ have similar structures with a core triple helical domain of 1014 amino acids composed of uninterrupted Gly-X-Y tripeptide repeats, where Gly is glycine and X and Y are often proline or lysine, flanked by propeptides at both N- and C-terminal ends. During and after translation, the three chains undergo extensive modification. Prolyl-4-hydroxylase converts virtually all Y-position proline residues to 4-hydroxyproline, an alteration that is essential for thermal stability of the assembled trimer. In the absence of this modification the trimer melts (that is the individual chains unfold from the stable triple helix, at about 27°C, whereas with full hydroxylation, the melting temperature is about 42°C). Some Y-position lysine residues within the triple helical domain are hydroxylated by the enzyme lysyl hydroxylase-1, and glucose and galactose groups added by glycosyltransferases (Fig. 31.1). Hydroxylation of these triple helical residues is part of the pathway to form

stable complex intermolecular crosslinks that provide the tensile strength in tissues. Most of these modifications are completed during translation and occur on the individual chains. If there is a delay in triple helix folding, the process can continue and the physical properties of the chains and molecules are altered and contribute to an osteogenesis imperfecta phenotype.

The three chains that form a [$\text{pro}\alpha 1(\text{I})_2 \text{pro}\alpha 2(\text{I})$] trimer interact through regions in the carboxyl-terminal propeptide of each chain. The full length chain must be maintained in an unfolded state while the carboxyl-terminal propeptides fold, associate, and then begin the process of triple helix formation. Propagation of the collagen triple helix requires a number of enzymes and molecular chaperones to ensure correct folding and trimerization. These include peptidyl disulfide isomerase (PDI), which also forms part of the prolyl 4-hydroxylase complex, and is likely to involve prolyl peptidyl cis-trans isomerase B (also known as cyclophilin B). This protein can act on its own, to assist in the folding around prolyl residues, such as those in the carboxyl-terminal propeptide adjacent to cysteine residues, and as part of a complex that includes two additional proteins, cartilage-related protein and prolyl 3-hydroxylase, to modify certain triple helical prolines. The function of this last process is not yet entirely clear but when the complex is missing, the propagation of the triple helix is altered and modification of the chains increased.

Disulphide bonds between the carboxyl-terminal region of the chains act to secure the three chains in a trimer, a process requiring protein-disulphide isomerase. Lysine residues outside the major triple helical domain of type I collagen, needed for the formation of mature intermolecular crosslinks, are hydroxylated by lysyl hydroxylase-2. These complex modifications, which are necessary for correct folding, assuring thermal stability of the triple helix and crosslink formation between collagen molecules once they are secreted into the matrix, need to take place in an orderly and timely sequence, and various chaperone proteins, including HSP47 and FKBP65, help regulate this process. Procollagen trimers are then transported via the Golgi network and packaged into membrane-bound organelles where lateral aggregation, the initial phase of fibril formation, occurs. As secretion occurs, the procollagen molecules are further processed into mature type I collagen molecules by proteolytic cleavage of the N- and C-terminal propeptides (by the enzymes ADAMTS-2 (a disintegrin and metalloproteinase with thrombospondin motifs 2) and BMP1 (bone morphogenetic protein 1), respectively). Finally, the trimers are assembled into collagen fibrils and fibres and anchored in those positions by intermolecular lysine-derived crosslinks in a process that is begun by modification of specific residues by lysyl oxidase. These cross-links are initially reducible, but as tissue maturation proceeds, they are converted into non-reducible compounds including hydroxylysylpyridinoline (derived from three hydroxylysine residues) and the less abundant lysylpyridinoline (derived from two hydroxylysine residues and one lysine residue). The latter (also known as deoxypyridinoline) is present in the body almost exclusively in bone tissue and dentine, where it accounts for 21% of the total mature cross-links.

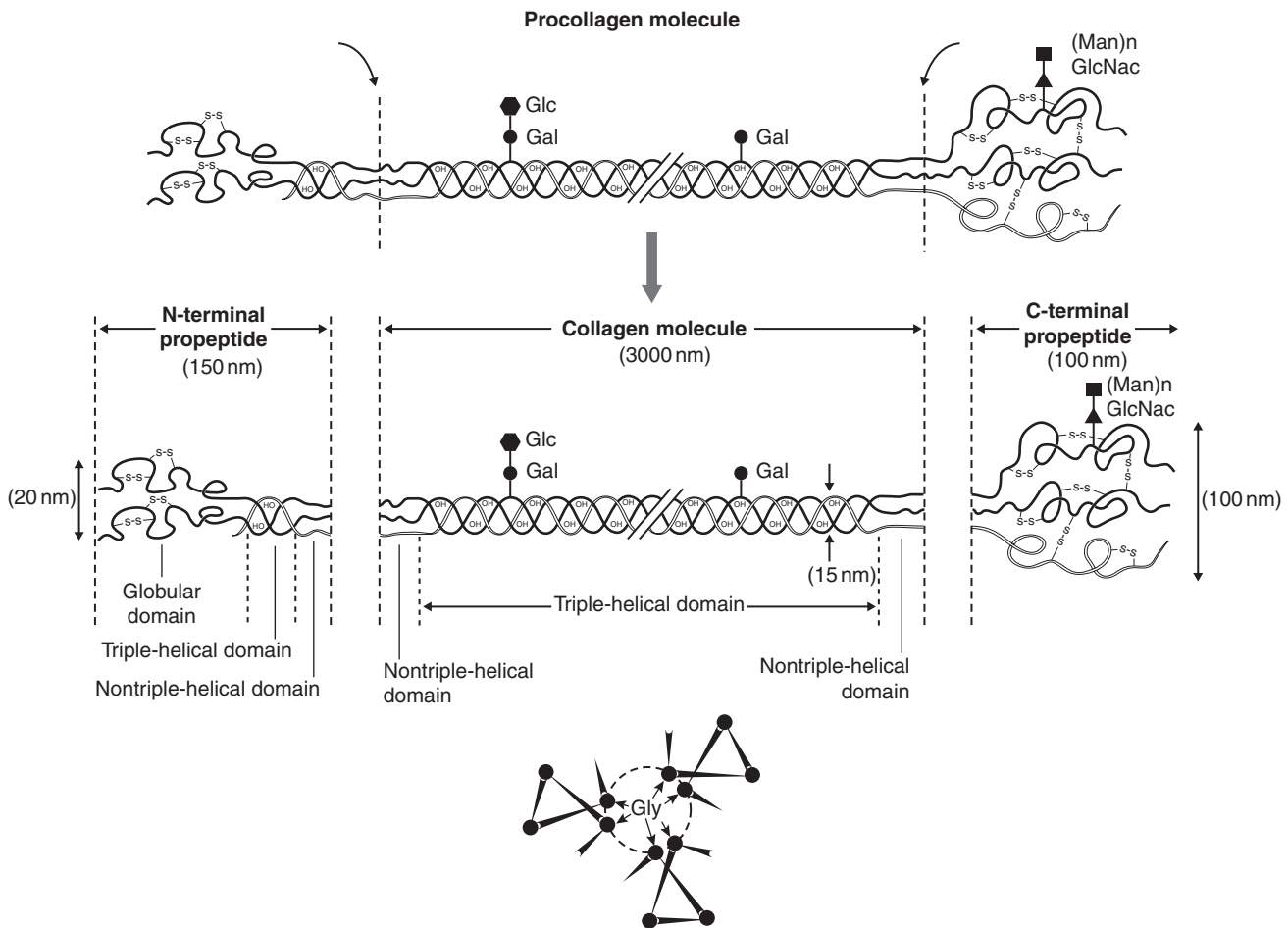


FIGURE 31.1 ■ Type I collagen synthesis. *Upper panel:* schematic representation of the structure of the procollagen molecule, showing cleavage of the amino and carboxy terminal peptides. Glc, glucose; Gal, galactose; Man, mannose; GlcNac, N-acetylglucosamine. *Lower panel:* view along the axis of the triple helical structure of the collagen backbone, with every third amino acid a glycine residue forming the inner circle.

Non-collagenous proteins

The non-collagenous proteins of bone (proteoglycans, glycosylated proteins, RGD (arg-gly-asp)-containing glycoproteins and γ -carboxylated proteins) account for only 10–15% of bone protein mass but, because of their very much smaller size, are as common as collagen in bone on a molar basis. Most are synthesized by osteoblasts, but a number are produced elsewhere in the body and arrive in bone via the circulation. The latter proteins, which include albumin and platelet-derived growth factor, are usually negatively charged and become bound to hydroxyapatite. The most abundant of the bone-derived non-collagenous proteins are osteocalcin, also known as bone Gla-protein, and osteonectin, both of which have a high affinity for bone mineral.

Osteocalcin is synthesized by osteoblasts as a 75-amino acid precursor known as pro-osteocalcin. The amino-terminal propeptide includes the substrate recognition site for the vitamin K-dependent γ -carboxylase enzyme which γ -carboxylates three glutamic acid residues (17, 21 and 24) in the central region of the osteocalcin molecule, following which the amino-terminal propeptide is cleaved, leaving the 49-residue osteocalcin molecule.

The γ -carboxyglutamic acid residues are strongly calcium binding and allow the molecule to bind to hydroxyapatite. The process of γ -carboxylation seems to be impaired in the elderly.

The physiological role of osteocalcin is uncertain, but it probably contributes to the regulation of bone remodelling, since deletion of the osteocalcin gene in mice leads to increased bone density. Its synthesis is regulated by 1,25-dihydroxyvitamin D (1,25(OH)₂D, calcitriol) and cortisol, and it is incorporated into bone matrix, being especially abundant in cortical bone.

Recently, its uncarboxylated form has been implicated in the regulation of intermediary metabolism, including fat mass, glucose tolerance and insulin secretion. The relevance to human physiology of these findings in genetically modified mice remains to be determined.

Other bone proteins

Several proteoglycans, macromolecules that contain acidic polysaccharide side chains, are present in bone. In the early stages of bone formation, versican, a dermatan sulphate proteoglycan, and the glycosaminoglycan, hyaluronan, are present and may act to delineate areas that are destined to become bone. Subsequently, two chondroitin

TABLE 31.1 Major glycoproteins containing the arginyl-glycyl-aspartate (RGD) cell-attachment motif in skeletal tissue

Protein	Proposed function
Thrombospondins	Cell attachment, bind heparin, collagens, thrombin, fibrinogen, plasminogen
Fibronectin	Binds to cells, collagen, fibrin, heparin, gelatin
Vitronectin	Cell attachment, binds collagen and plasminogen
Fibrillin	Regulation of elastic fibre formation
Osteopontin	Binds to cells, inhibits mineralization, regulates cell proliferation and tissue repair
Bone sialoprotein	Binds cells, may inhibit mineralization
BAG-75	Binds calcium, ?cell attachment
Dentin matrix phosphoprotein 1	Osteocyte-derived, regulates osteocyte maturation and FGF23 production

sulfate proteoglycans, decorin and biglycan, are expressed and may influence cellular proliferation and differentiation in the newly forming bone. Fibromodulin, osteoglycin and osteoadherin are other proteoglycans that are expressed in bone matrix, but their functions are less well characterized.

The major glycoproteins present in bone are alkaline phosphatase (ALP), which is highly expressed on the osteoblast cell surface, and osteonectin, a phosphorylated glycoprotein that is also present in several other tissues that undergo rapid remodelling. In bone, osteonectin may regulate both cellular activity and matrix mineralization.

At least eight non-collagenous glycoproteins expressed in bone contain the arginyl-glycyl-aspartate (RGD) tripeptide as the consensus cell-attachment sequence (Table 31.1). Some are specific to bone (e.g. bone sialoprotein), but most are more ubiquitously expressed (e.g. fibronectin, osteopontin, thrombospondins, vitronectin). Although the role(s) of each individual protein in skeletal physiology is(are) uncertain, it seems very likely that they collectively serve to promote attachment of bone cells to bone matrix (and perhaps thereby regulate their function). The components of the matrix–cell interactions are clearly dependent upon the maturational stage of the remodelled bone. Three matrix non-collagenous proteins, matrix Gla-protein, protein S and osteocalcin (bone Gla-protein), undergo post-translational γ -carboxylation, a process that enhances calcium binding. These proteins may function physiologically as inhibitors of mineral deposition.

Cellular elements of bone

Osteoblasts

Osteoblasts are the cells responsible for the synthesis and maintenance of bone matrix. They are derived from pluripotent mesenchymal progenitor cells that also give rise

to adipocytes, chondrocytes and myocytes. Expression of the transcription factors *cbfa1* and *osterix* is critical for commitment to the osteoblast lineage, while expression of peroxisome proliferation activating receptor- γ (PPAR- γ) favours adipocytic over osteoblastic differentiation.

A key molecular event in osteoblast proliferation, and perhaps survival, is signalling through the wnt-low density lipoprotein-related receptor 5 (LRP5)- β -catenin pathway. In this system, secreted wnt proteins bind to and activate LRP5 on the cell surface, which in turn activates a series of intracellular events that result in translocation of cytoplasmic β -catenin to the nucleus, where it activates transcription of genes critical for osteoblast growth and function. The pathway is further regulated by secreted proteins that inhibit wnt-induced activation of LRP5, such as sclerostin, Dickkopf (DKK) and frizzled. Thus, humans carrying mutations of components of this signalling pathway that decrease nuclear translocation of β -catenin have decreased bone mass and fractures as a result of impaired bone formation (e.g. *WNT1* and most *LRP5* mutations), while constitutive activation of the pathway (e.g. *LRP5* mutations affecting one particular extracellular domain or *SOST* mutations) increases bone mass. Other LRP family members are also important in bone physiology. Mutations affecting one particular region of the extracellular domain of LRP4 result in impairment of the ability of sclerostin to inhibit bone formation, and thereby cause increased bone mass (sclerosteosis type 2).

Mature osteoblasts are metabolically very active and intimately involved in bone formation. They synthesize many of the matrix proteins referred to above. They have very high levels of alkaline phosphatase (ALP) activity associated with their cell membranes. This enzyme is necessary for normal bone mineralization to take place. Osteoblasts express receptors for parathyroid hormone (PTH) and $1,25(\text{OH})_2\text{D}$, as well as a number of cytokines, growth factors and sex hormones. Osteoblasts also play a key role in regulating osteoclastic bone resorption, ensuring that the components of bone remodelling, formation and resorption are coupled. A critical mechanism by which osteoblasts regulate bone resorption is via the RANK–RANKL–OPG system (see below).

Osteocytes

Osteocytes and bone lining cells are terminally differentiated osteoblasts derived from mature osteoblasts that are no longer involved in active bone formation, and have become entrapped within the canaliculi of the bone matrix that they have produced. This process of osteocytogenesis is probably an active one, requiring cleavage of matrix proteins: mutations of the metalloproteinase MT1-MMP lead to osteocytes with fewer and shorter dendritic processes. As osteocytogenesis occurs, the osteoblast changes from a cuboidal cell to a typical osteocytic morphology, characterized by a stellate appearance and numerous cytoplasmic projections that permit the cell to communicate with cells on the bone surface and other osteocytes living within adjacent bone. Concurrently, the genetic signature of the cell changes from a typically osteoblastic one to that which reflects a mature osteocyte, in particular the expression

of *DMP1*, *SOST* and *FGF23*. Osteocytes play a key role in the regulation of bone remodelling. At sites of skeletal microdamage, osteocytes undergo apoptosis and release apoptotic bodies expressing RANKL (see below), promoting the recruitment of osteoclasts to the area to initiate remodelling to allow repair of the damaged bone. Osteocytes also produce and release sclerostin, an endogenous inhibitor of osteoblast differentiation and function that acts to inhibit wnt-induced activation of LRP5. They also act as mechanosensory cells in bone, perhaps in tandem with osteoblasts lining the bone surface, with which they communicate. Finally, osteocytes are critically important for bone mineralization, by regulating phosphate metabolism at a whole organism level, principally by production of the phosphaturic hormone fibroblast growth factor 23 (*FGF23*), a process that involves the osteocyte products phosphate-regulating endopeptidase (*PHEX*) and dentine matrix phosphoprotein 1 (*DMP1*).

Osteoclasts

Osteoclasts are the cells that resorb bone. They are derived from cells of the monocyte/macrophage series. They are multinucleated and are usually found singly or in small numbers within resorption lacunae of their own making. They have large numbers of mitochondria and lysosomes. Activated osteoclasts possess a 'ruffled' border, adjacent to the bone being resorbed. The ruffled border is delimited by a 'sealing zone' at which the osteoclast is attached to the underlying bone. Bone resorption occurs as a result of the secretion of protons and proteolytic enzymes into the space between the ruffled border and the bone surface. Among these enzymes are tartrate-resistant acid phosphatase (a useful histological marker of osteoclasts) and cathepsin K.

Osteoclast development and function are regulated by osteoblast and osteocyte-derived cytokines, in particular osteoprotegerin (*OPG*) and receptor activator of nuclear factor- κ B ligand (*RANKL*), which are members of the tumour necrosis factor (*TNF*) receptor superfamily. *RANKL* expressed on the osteoblast and osteocyte surface binds to, and activates its receptor – receptor activator of nuclear factor- κ B (*RANK*) – which is expressed on the surface of osteoclast precursors. *RANKL* is also shed from the cell surface, and this soluble form of the protein may confer osteoclastogenic activity distant from the cell of origin, in particular the osteocyte. Interaction between *RANKL* and *RANK* promotes osteoclastogenesis. Osteoprotegerin is a secreted osteoblast product that acts as a decoy receptor for *RANKL*, and thereby functions as an endogenous inhibitor of osteoclastogenesis. Thus, the relative levels of expression by the osteoblast of *RANKL* and *OPG* determine the rate at which osteoclastogenesis occurs. Many of the systemic factors that alter bone resorption, such as *PTH* and $1,25(\text{OH})_2\text{D}$ (which increase osteoclast function) and sex hormones (which decrease it), do so by changing osteoblastic production of *RANKL* and/or *OPG*. An exception is calcitonin, whose receptor is expressed on osteoclasts, and which therefore inhibits bone resorption directly.

A number of other critical cellular and molecular events are now recognized to underpin osteoclastic

bone resorption. The intracellular tyrosine kinase *c-src* regulates osteoclast motility and aspects of cytoskeletal function that are critical to formation of a functional resorption space; the $\alpha_v\beta_3$ integrin (vitronectin receptor) plays a key role in osteoclast attachment to bone; the cysteine protease cathepsin K and the type 7 transmembrane chloride channel (*ClCN7*) are each required for proteolytic degradation of bone matrix.

Bone remodelling and its regulation

In mature bone, there is a continuous process of removal and replacement of pockets of old bone. This is termed *bone remodelling* and is shown in schematic form in Figure 31.2. The process occurs at discrete sites on the bone surface and, in the mature skeleton, is probably an adaptive response to skeletal microdamage. Bone remodelling is achieved by teams of osteoclasts and osteoblasts acting within the anatomical structure known as the basic multicellular unit (*BMU*). Osteocytes play a critical role in initiating bone remodelling, and probably also in regulating the coordination of formation and resorption within the *BMU*. Thus, apoptotic osteocytes in an area of skeletal microdamage produce *RANKL* that directs osteoclastogenesis and initiation of bone remodelling to the damaged bone. At that site, the osteoclasts excavate a resorption pit, the depth of which is determined by the number and activity of the osteoclasts present. The motile osteoclasts are then replaced by osteoblasts, which proceed to fill the resorption pit with bone matrix. Mineralization of this new bone then occurs, lagging behind matrix synthesis by a period of about three weeks. Once the resorption pit has been filled, the osteoblasts either undergo apoptosis, return to their quiescent state as bone lining cells or survive as osteocytes within the bone lacunae.

An increase in bone turnover is characterized by a faster bone remodelling cycle initiation rate, rather than an increase in the duration of the cycle. The histological consequence of such an increase is that a greater proportion of bone surface is involved in remodelling at any one time. The time required to complete a full cycle of resorption and formation (σ) is in the order of 6–12 months. Inhibiting bone resorption, for example with pharmaceutical agents, rapidly reduces the initiation rate of new remodelling cycles, but those cycles already initiated go to completion. Thus there is a period ('a transient') of duration σ , when the bone formation rate exceeds the resorption rate. This ends when all those cycles initiated before the inhibition of bone resorption took place have completed their formation period. The bone formation rate at this stage will have fallen to match the new, lower, resorption rate.

Bone resorption and formation are thus very closely linked in normal bone and also in many metabolic bone disorders. As outlined earlier, the osteoclasts and osteoblasts clearly 'communicate' with each other. Osteoblastic regulation of osteoclast function is mediated in large part by the *RANK*–*RANKL*–*OPG* system, but the nature of the osteoclast-derived signal(s) that recruits and activates osteoblasts is not yet fully characterized. Osteocyte-derived sclerostin, an inhibitor of osteoblast differentiation, probably contributes to the regulation

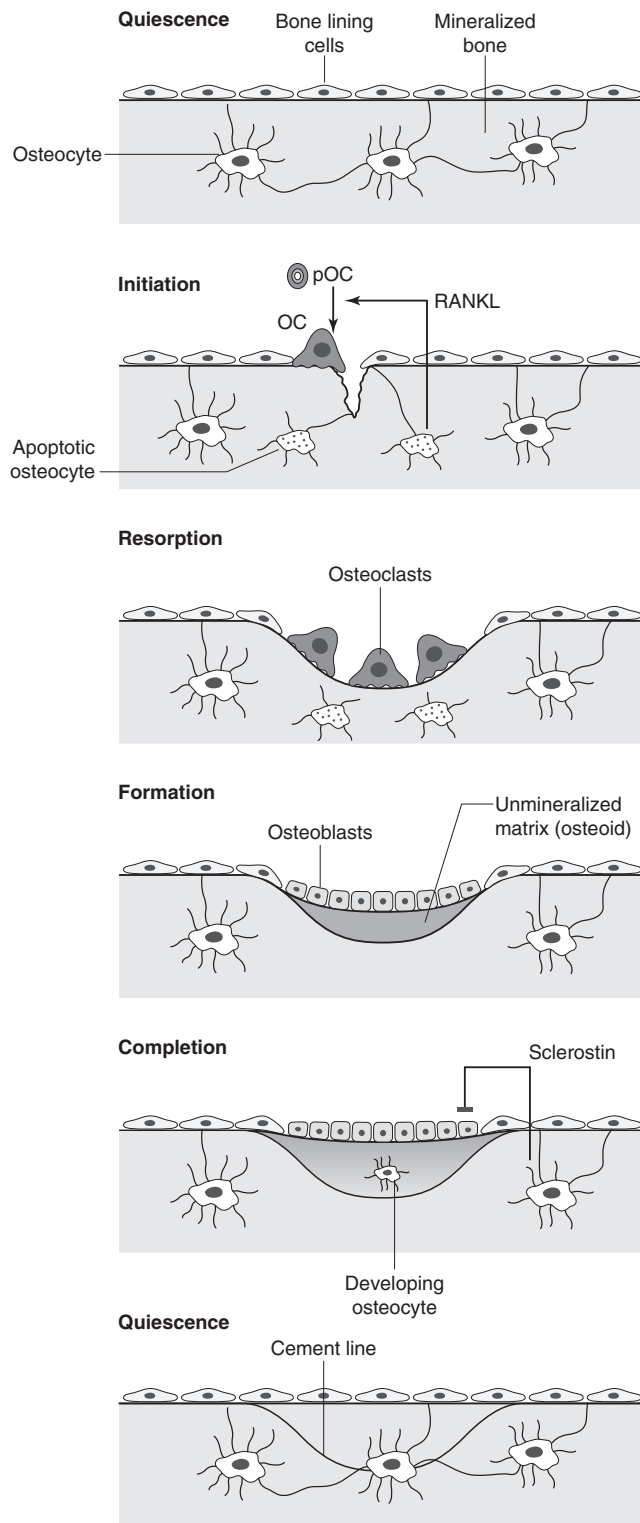


FIGURE 31.2 ■ The bone remodelling cycle. Bone remodelling is initiated in response to skeletal microdamage, sensed by adjacent osteocytes which undergo apoptosis and produce the osteoclastogenic cytokine, receptor activator of nuclear factor- κ B ligand (RANKL). Osteoclastic resorption of the damaged bone ensues, after which mature osteoblasts progressively fill the resorption cavity with osteoid, and coordinate its subsequent mineralization. Some of the osteoblasts develop into osteocytes within the lacunae of the newly formed bone. Sclerostin is produced by osteocytes adjacent to the remodelling space: it signals inhibition of bone formation and a return to a quiescent state. In skeletal homeostasis, the amount of new bone formed equals that resorbed during each remodelling cycle. OC, osteoclast; pOC, pre-osteoclast. (Courtesy of Dr Andrew Grey.)

of the bone formation component of the remodelling cycle. Matrix-associated growth factors that are released from resorbed bone, such as transforming growth factor β (TGF β), bone morphogenetic proteins (BMPs) and platelet-derived growth factor, may also constitute part of this signal. Finally, genetic evidence suggests that communication from osteoclasts to osteoblasts occurs via members of the ephrin (Eph) family of signalling proteins, specifically osteoclast-expressed ligand ephrin-B2 signalling through the osteoblast-expressed receptor EphB4 to activate osteoblast differentiation and bone formation.

During growth, the amount of bone deposited exceeds that which is removed and there are also changes in the size and configuration of bones (modelling). In senescence, because of factors including sex hormone deficiency, reduced physical activity and various endocrine and inflammatory diseases, the balance of formation and resorption is reversed and bone is lost. Age-related bone loss affects both cortical and trabecular bone, but early after the menopause, bone loss is particularly rapid from the axial skeleton, which has a greater proportion of trabecular bone.

A large number of factors have been found to influence bone remodelling (discussed in Chapter 6).

Biochemical markers of bone turnover

Bone turnover is a term referring to both osteoblastic bone formation and osteoclastic bone resorption. These processes can be assessed by bone histomorphometry, skeletal scintigraphy and by analysing the kinetics of intravenously administered isotopes of calcium. There are numerous biochemical markers of bone turnover, representing either products of bone cells or degradation products from the breakdown of bone collagen (Box 31.1). The number of potential markers available has increased considerably in recent years. In a small number of rare conditions (e.g. hypophosphatasia, osteopetrosis, Bruck syndrome and mild osteogenesis imperfecta), a specific pattern of bone turnover markers can be helpful in diagnosis. However, the main use of markers is to establish whether a high bone turnover state exists. High bone turnover, with its various adverse consequences, is a feature of many of the more common metabolic bone disorders, and a key target of currently available therapies.

Markers of bone formation

Alkaline phosphatase. Alkaline phosphatase is widespread in human tissues. Enzymes with this activity are coded for by genes at three separate loci, their respective products being referred to as intestinal, placental and bone/liver/kidney (or tissue non-specific) ALP. The latter gene is located on chromosome 1p36. The gene transcript has a number of glycosylation sites, and the bone, liver and kidney isoenzymes differ in their carbohydrate side chains.

The major limitation of total ALP as an index of bone formation is its multiple tissues of origin and hence its lack of sensitivity at low values. In health, plasma ALP activity is principally derived from the liver and osteoblasts in approximately equal proportions. If the activity of other

BOX 31.1 Biochemical markers of bone turnover
Formation markers
Blood

- Alkaline phosphatase (ALP) (total or bone specific)
- Osteocalcin (OC)
- Procollagen 1 carboxy-terminal extension peptide (P1CP)
- Procollagen 1 amino-terminal extension peptide (P1NP)*

Resorption markers
Blood

- Cross-linked C-telopeptide of type I collagen (ICTP)
- N-telopeptide of collagen cross-links (NTX)
- C-telopeptide of collagen cross-links (CTX)*
- Tartrate-resistant acid phosphatase (TRAP)
- Cathepsin K

Urine

- Hydroxyproline
- Hydroxylysine
- Pyridinoline (total and free)
- Deoxypyridinoline (total and free)
- N-telopeptide of collagen cross-links (NTX)
- C-telopeptide of collagen cross-links (CTX)

*P1NP and CTX have been endorsed by the International Osteoporosis Foundation for use in clinical trials of osteoporosis treatments.

hepatic canalicular cell enzymes is normal, then elevated total ALP activity usually reflects osteoblastic activity (or, more precisely, the number of active osteoblasts). If uncertainty persists in determining the origin of increased plasma ALP, then isoenzyme studies may be undertaken. The bone and liver isoenzymes have differing stabilities to heating (the bone fraction being more sensitive to heat inactivation) and different mobilities on electrophoresis. Immunoassays for bone-specific ALP use isoform-specific monoclonal antibodies; even so, there is up to 20% cross-reactivity with the liver isoform in these assays. Thus very high plasma activities of liver ALP, as occur in hepatobiliary disease, can still interfere with the measurement of bone-specific ALP. The activity of the bone isoenzyme has a diurnal variation in plasma, with peak values occurring around midnight and being some 25% higher than early morning values. Although the half-life in plasma of the bone isoenzyme is very short (~2 days), in metabolic bone disorders changes in ALP activity reflect slower processes – the acceleration or deceleration of the remodelling cycle – so the rate of change of plasma ALP is also much slower.

The marked elevation in plasma ALP activity seen in childhood and adolescence is due to an increase in the bone fraction. The plasma ALP activity closely parallels the growth velocity curve (Fig. 31.3). A significant (40–50%) rise in plasma ALP activity occurs at the menopause because of accelerated bone turnover. Plasma activity of the enzyme can also be increased after major fractures and reflects the increased cellular activity associated with healing. After hip fracture, for example, ALP activity reaches a peak at around four weeks, with values approximately double those on admission to hospital. Total plasma ALP activity is increased during pregnancy because of the appearance of the placental isoenzyme.

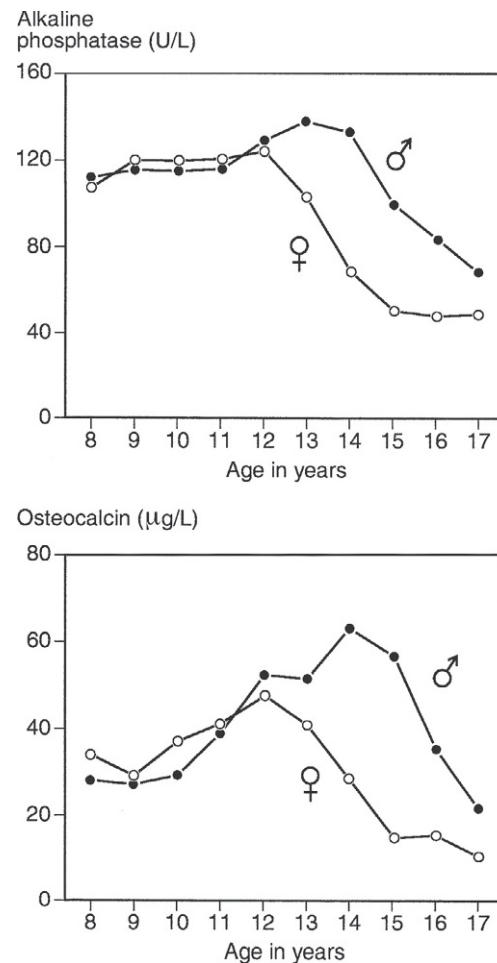


FIGURE 31.3 ■ Bone turnover markers during growth. Mean values of plasma alkaline phosphatase (ALP) and osteocalcin in males (closed circle) and females (open circle) through childhood and adolescence. Peak values are reached at an average age of 12 years in girls and 14 in boys, coinciding with peak growth velocity. As growth velocity falls, both these indices decline toward adult values. (Data from Round JM 1973 Plasma calcium, magnesium, phosphorus and alkaline phosphatase concentrations in normal British school children. *British Medical Journal* 3: 173–140, and Johansen JS, Riis BJ, Podenphant J, Skakkeboek NE, Christiansen C 1987 Plasma bone Gla protein: a specific marker of bone formation. In: Christiansen C, Johansen JS, Riis BJ (eds). *Osteoporosis*, vol 2. Copenhagen: Osteopress ApS: 677–681.)

The entity of *familial benign hyperphosphatasia* is characterized by asymptomatic high ALP activities (predominantly of intestinal isoform) and dominant inheritance, but with no clinical evidence of bone disease. *Benign transient hyperphosphatasemia* is a condition characterized by high concentrations of both bone and liver isoenzymes, first identified in children but now recognized in adults (see Chapter 13).

There are a number of causes of low ALP activities (*hypophosphatasemia*), including vitamin C deficiency, hypothyroidism, starvation, zinc or magnesium deficiency, Cushing syndrome, Wilson disease and the genetic disorder, hypophosphatasia (discussed below).

Osteocalcin. Osteocalcin is produced by osteoblasts, and is generally regarded as a marker of bone formation, but it seems to be involved in the process of mineralization

rather than matrix production. Plasma concentrations of intact osteocalcin are markers of bone formation, and circulating concentrations generally correlate well with histological measures of bone formation rate. Around 25% of circulating osteocalcin consists of intact osteocalcin, the remaining immunoreactivity comprising N-terminal, mid-region, mid-region-C-terminal and C-terminal fragments. When bone is resorbed, osteocalcin fragments are released, and the plasma concentrations of these fragments may thus reflect bone resorption. These fragments are cleared predominantly by the kidneys, so there is considerable immunological heterogeneity in plasma osteocalcin in patients with renal impairment. These factors may limit the usefulness of osteocalcin as a marker of bone formation.

Osteocalcin has a particular advantage in having greater sensitivity than ALP in the detection of low rates of bone formation. As with plasma ALP activity, plasma osteocalcin and growth rate are closely correlated (see Fig. 31.3), and there is a two-fold rise in plasma osteocalcin concentration at the menopause. Plasma osteocalcin concentrations fall markedly during pregnancy. Cord blood concentrations are two to three times higher than adult values. The half-life of intact osteocalcin in the circulation is about 5 min.

Plasma osteocalcin concentrations exhibit a diurnal variation with the peak in the early morning being about 15% higher than the nadir value, which occurs in the afternoon. Concentrations may also be depressed (by about 10%) by alcohol intake and are markedly suppressed (by about 50%) by glucocorticoids. Plasma concentrations rise when osteoblasts are stimulated by the administration of 1,25(OH)₂D.

Procollagen 1 extension peptides. Antibodies have been raised to both the soluble C-terminal and N-terminal propeptides of procollagen, which appear in the circulation after cleavage from the newly formed collagen molecule. Their concentrations in the plasma thus reflect osteoblastic collagen synthesis. As they are not filtered by the kidney, variations in renal function should not affect their plasma concentrations. Degradation of the proteins takes place in hepatic endothelial cells (through the mannose 6-phosphate receptor in the case of procollagen 1 carboxy-terminal extension peptide (P1CP), and by scavenger receptors in the case of procollagen 1 amino-terminal extension peptide (P1NP)).

Procollagen 1 carboxy-terminal extension peptide is usually measured by enzyme linked immunosorbent assay (ELISA). It has a diurnal variation, with peak concentrations occurring at 04.00–08.00 h, and the nadir in the afternoon. There may be some contribution to circulating P1CP from non-osseous sources.

Procollagen 1 amino-terminal extension peptide has been more widely studied. It appears to be a more dynamic marker than P1CP or other formation markers. The peptide is stable, and the available assays are precise and responsive to common interventions in metabolic bone diseases. It is now regarded as the marker of choice to assess osteoblast activity. Of the various assays available for P1NP, some measure the intact propeptide (intact P1NP) while others (total P1NP) also detect a smaller antigen in serum. In many clinical situations, these assays give similar information, but renal insufficiency increases the

concentrations of the smaller antigen, influencing both the apparent concentration of P1NP and assay calibration.

Markers of bone resorption

Hydroxyproline. Hydroxyproline is a major component of fibrillar collagen of all types, comprising ~14% of the total amino and imino acid content. It is produced by the post-translational modification of proline by the enzyme 4-prolyl hydroxylase. In plasma, hydroxyproline exists in protein-bound, peptide-bound and free forms. Most hydroxyproline is oxidized in the liver but a small proportion (~10%) is excreted in the urine. Approximately 90% of collagen-derived hydroxyproline excretion is in the form of peptides resulting from collagen degradation, with the remainder in the form of fragments of the N- and C-terminal procollagen peptides cleaved during collagen synthesis. When bone resorption is increased, urinary hydroxyproline excretion rises, because of the increased rate of catabolism of type 1 bone collagen. When bone turnover is low, urinary hydroxyproline is not a reliable index of bone resorption, since the other sources (for example C1q complement component) can contribute up to 50% of the total urinary hydroxyproline.

Like all urinary collagen breakdown products, hydroxyproline excretion can be measured in a 24 h urine collection or, more conveniently, in fasting, second voided morning samples, expressed as a ratio to the creatinine concentration. Dietary sources of hydroxyproline (e.g. gelatin) need to be eliminated immediately before and during the collection. The major limitations of hydroxyproline estimation as an index of bone resorption are interference from sources other than type 1 collagen, its lack of sensitivity at low concentrations, its day-to-day variability and its partial origin from bone formation.

Glycosylated hydroxylysine. Hydroxylysine – another amino acid unique to collagen-like peptides – is produced by the post-translational modification of lysine residues. These amino acids can then undergo further modification by glycosylation – giving rise to galactosyl-hydroxylysine and glucosylgalactosyl-hydroxylysine. The former predominates in the type 1 collagen of bone. Galactosyl-hydroxylysine is not significantly metabolized before excretion in the urine (10% in free form, 90% peptide bound), and the urinary concentrations are not affected by diet, so it is a good marker of bone resorption. It can be measured by HPLC, but the assay is technically demanding.

Collagen cross-links. The amino acids hydroxylysylpyridinoline and lysylpyridinoline (also known as pyridinoline and deoxypyridinoline, respectively) are created by the covalent cross-linking that takes place when mature collagen molecules stabilize after forming the triple helical structure (Fig. 31.4). Since the cross-links are formed only in extracellular collagen fibrils, they appear in the circulation and the urine only as degradation products of mature matrix and do not reflect new collagen synthesis. In contrast to hydroxyproline, neither pyridinoline nor deoxypyridinoline occur in the collagen of normal skin. Pyridinoline is present in other connective tissues, notably cartilage and tendon, whereas the

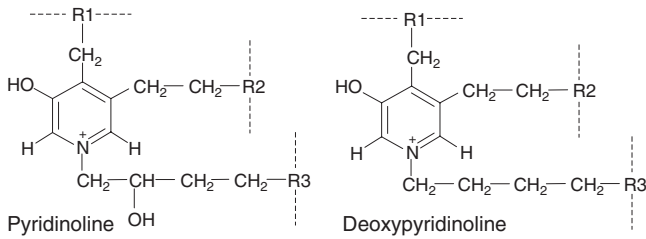


FIGURE 31.4 ■ Molecular structure of pyridinoline cross-links. R1 and R2 are telopeptide sequences and R3 a helical fragment sequence; for free crosslinks, R1, R2 and R3 are: $-\text{CH}(\text{COOH})\text{NH}_2$.

less abundant deoxypyridinoline is more specific to bone and dentine.

In urine, both pyridinoline and deoxypyridinoline are present as both free amino acid derivatives (~40%) and as oligopeptide-bound fractions (~60%). The free forms can be measured directly, but the conjugated form has to be hydrolysed before assay. After acid hydrolysis, total urinary pyridinoline and deoxypyridinoline can be measured by HPLC, but ELISA immunoassays for the free amino acids or peptides containing the cross-links are now available.

The pyridinoline:deoxypyridinoline molar ratio in urine is similar to that found in bone (7:2), and the urinary excretion rate of the cross-links agrees well with radioisotopic and histomorphometric measures of bone resorption. The cross-links do not undergo further metabolism once released, and their rate of excretion is not influenced by dietary intake or moderate renal impairment. Random urine measurements (expressed relative to creatinine concentration) correlate closely with 24h estimates. The elderly tend to have relatively greater excretion of the larger peptide-bound forms, and a correspondingly

smaller proportion of the free pyridinoline and low molecular weight peptide forms.

Urinary pyridinoline and deoxypyridinoline concentrations appear to be good markers of bone resorption, and more sensitive than hydroxyproline in distinguishing small increases in bone resorption from normal. For example, when normal premenopausal and postmenopausal women are compared, hydroxyproline excretion in the latter group is about 50% above that of premenopausal women, whereas for urinary pyridinoline and deoxypyridinoline, excretion is increased by 100%.

Collagen telopeptides. Collagen cross-links form mainly between the short non-helical peptides at both ends of the collagen molecule (the N- and C-terminal telopeptides), which link via pyridinium or pyrrole compounds to the helical portion of an adjacent collagen molecule (Fig. 31.5). During collagen breakdown, N- and C-terminal telopeptide fragments, still attached by the pyridinium cross-link to the helical fragments of the nearby molecule, are released into the circulation and are cleared by the kidneys. Within the C-telopeptides of type I collagen is an aspartyl-glycine motif that can undergo β -isomerization to isoaspartyl-glycine. This process is believed to be linked to protein ageing, with younger bone having a greater proportion of the non-isomerized α -C-telopeptides, and older bone having a greater proportion in the β -isomerized C-telopeptide form. Various immunoassays have been developed to detect epitopes specific to particular fragments.

The urinary N-telopeptide (NTX) assay is an ELISA based on a monoclonal antibody that recognizes a conformational motif in the $\alpha 2$ -chain that is linked to a pyridinium cross-link. Although it does not recognize the pyridinoline or deoxypyridinoline cross-link itself,

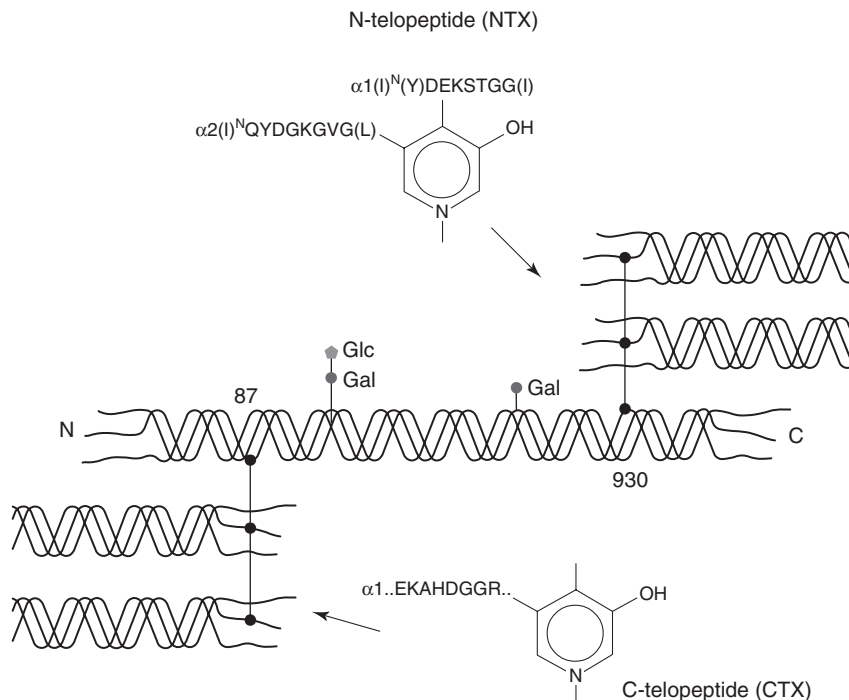


FIGURE 31.5 ■ Intermolecular cross-linking in type I collagen and the location of the peptide sequences of the N-terminal (NTX) and C-terminal (CTX) immunoassays.

it is specific for the cross-linking site of bone collagen. The assay requires no hydrolysis. The measurement is usually performed on a second void, morning urine sample, and the results are expressed as bone collagen equivalent, corrected for urinary creatinine concentration. A serum assay has also been developed, which correlates well with urinary concentrations.

Immunoassays for the C-terminal telopeptide cross-linked region of type I collagen (CTX) have also been developed. The monoclonal antibody recognizes an octapeptide sequence of the $\alpha 1$ chain containing the lysine in the C-terminal telopeptide that is involved in intermolecular cross-linking. Immunoassays that are specific for either α -CTX or β -CTX have been developed, and can measure these fragments in both serum and urine. Serum β -CTX is generally regarded as the marker of choice for assessing bone resorption. It should be measured fasting at 08.00–09.00 h, since it shows diurnal variation and is suppressed by food.

Tartrate-resistant acid phosphatase. The acid phosphatases are lysosomal enzymes that hydrolyse phosphomonoesters at low pH, releasing phosphoric acid. They are produced in a number of tissues, including bone, prostate, platelets and erythrocytes. Five isoenzymes are detectable in plasma. The different acid phosphatase isoenzymes differ in tissue of origin, chromosomal origin, molecular size and electrophoretic mobility. Based on the latter, the plasma isoenzymes have been classified as types 1–5. According to their sensitivity to inhibition by L(+)-tartrate, the isoenzymes are further classified as tartrate sensitive (TSAP) and tartrate resistant (TRAP).

Prostatic acid phosphatase (type 2b) activity is typically raised in plasma in carcinoma of the prostate. Type 5 acid phosphatase is produced by macrophages (TRAP5a) and osteoclasts (TRAP5b). TRAP5a activity in plasma is elevated in Gaucher disease and hairy cell leukaemia.

Tartrate resistant 5b differs from TRAP5a by not having sialic acid residues. It can be measured in serum by a variety of kinetic assays or by immunoassay. Measurement has some value in the diagnosis of osteoclast-rich forms of osteopetrosis. In most forms of this disorder, the number of osteoclasts in bone is actually increased – although they are ineffective at resorbing bone – and plasma activities of TRAP5b are increased. In most other bone disorders, TRAP5b plasma concentrations correlate well with other indices of bone resorption and with histological measurements of osteoclastic activity. Unlike other resorption indices, TRAP5b reflects osteoclast number and/or activity, rather than matrix breakdown. It may find a particular role in situations where these are dissociated, such as in examining the actions of drugs that inhibit osteoclast activity without reducing their number (e.g. cathepsin K inhibitors). Plasma acid phosphatase concentrations are also increased in bisphosphonate-induced osteosclerosis.

New markers. It is now possible to measure circulating cathepsin K concentrations. Preliminary reports indicate that these are higher in subjects with fractures, and are related to other turnover markers. Assays are also available for circulating osteoprotegerin, soluble RANKL, and sclerostin. These factors all act locally in bone, and the relevance

of circulating concentrations to judging their bioactivity in the bone micro-environment remains to be determined. These new markers are used mainly for research purposes at present, and their clinical utility is uncertain.

Variation in bone turnover markers. Many of the markers described are now widely used in research, and are finding a place in the management of some metabolic bone diseases. In general, they correlate well with assessments of turnover based on histomorphometry or calcium kinetic studies. The newer assays are more specific to bone and, in some cases, are more sensitive to perturbations, but their usefulness is limited by the day-to-day biological variability (as well as analytical imprecision). Stability is also an important issue for some analytes. In general, plasma markers of bone resorption show less day-to-day variability than urinary markers. Urinary resorption markers are usually measured on fasting second void urine samples to minimize the effect of diet and results expressed as a ratio to creatinine concentration. Thus, artefact can arise if muscle mass is significantly altered from normal.

Most markers of bone turnover show significant diurnal variation with highest values in the early morning hours and lowest values during the afternoon or night. The amplitude of this variability is around 15–30%. Therefore, resorption markers are best assessed in the early morning, after overnight fasting.

Bone turnover may also vary with the menstrual cycle: formation markers are higher in the luteal phase, whereas resorption markers are higher during the follicular phase. The amplitude of these changes is ~10%. After the menopause, there is an increase in bone turnover – resorption markers increase by 30–50% within 12 months of the menopause, and there are similar, but later, increases in formation markers. A less marked increase in bone turnover markers is also seen in men after the age of 50 years. Compared with adult values, all turnover markers are much higher during childhood and puberty, because the skeleton is growing and modelling, as well as remodelling (see below). It is thus important that age-appropriate reference ranges are used for all these assays.

Bone markers increase by ~50% two to four weeks after a fracture, and can take six to nine months to return to normal.

OSTEOPOROSIS

Osteoporosis is a reduction in bone density to the point that fractures occur as a result of the minor trauma that is a normal part of everyday life. At the microscopic level, osteoporotic bone is normally mineralized but reduced in volume. There is a generalized thinning of trabecular elements with the total loss of some trabeculae, greatly reducing the strength of bone (Fig. 31.6) and a reduction in cortical width in long bones. This loss of the normal microarchitecture emphasizes the need for prevention of bone loss, since restoration of the lost structure is unlikely to be achievable with pharmacological agents.

Osteoporosis can also be defined using bone mineral density (BMD) criteria. Thus, osteoporosis is defined as a BMD T-score less than –2.5, a BMD that is more

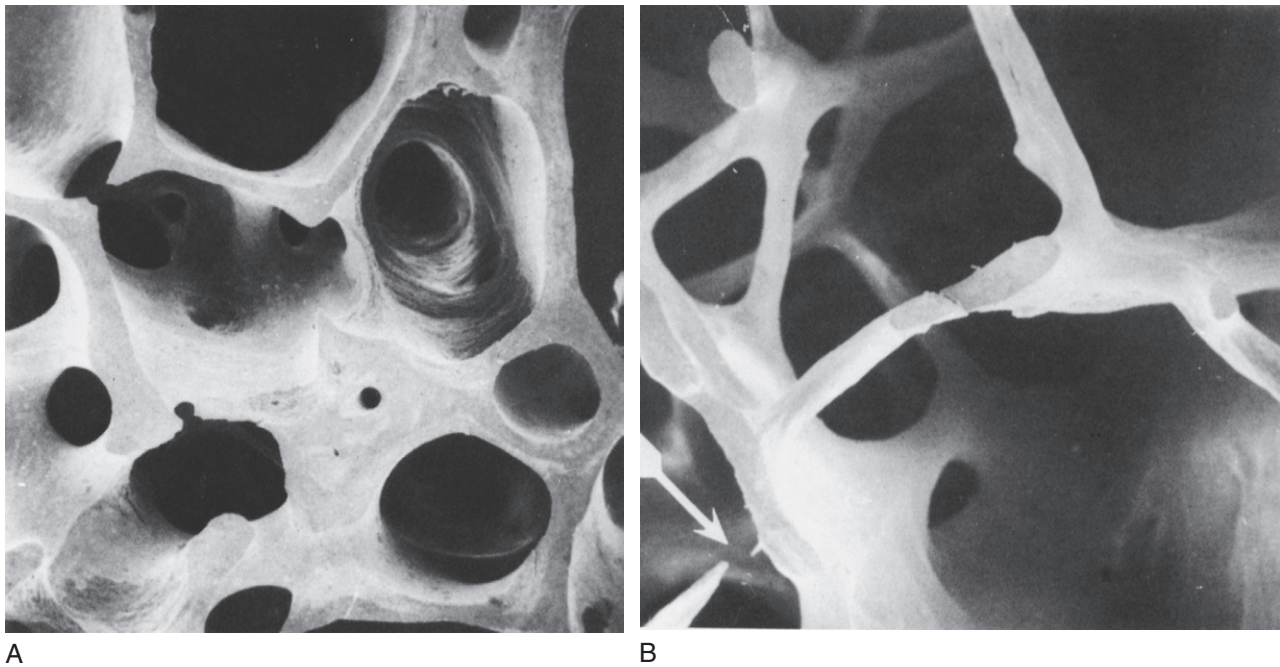


FIGURE 31.6 ■ Low-power scanning electron micrographs of iliac crest biopsies at the same magnification. (A) Taken from a healthy 47-year-old woman; (B) from an osteoporotic woman of 75 years. In comparison with (A), there has been thinning of the trabeculae with loss of some trabecular elements and discontinuity (arrowed) of some of those which remain. (From Dempster D W et al. 1986 A simple method of correlative light and scanning electron microscopy of human iliac crest bone biopsies. *Journal of Bone and Mineral Research* 1:21–25, with permission.)

than 2.5 standard deviations below the mean value in the young healthy population. Using this diagnostic criterion, ~50% of women have osteoporosis by the age of 80 years. Osteopenia is defined as a BMD T-score that is 1–2.5 standard deviations below the mean value in the young healthy population. Low bone mineral density is an important predictor of future fractures. Loss of BMD is sometimes pathological, but more frequently occurs as a normal part of ageing. With age, the efficiency with which osteoblasts refill resorption cavities is reduced, and there is also an increase in bone resorption associated with hypogonadism.

The important clinical consequence of osteoporosis is fracture. Common osteoporotic fractures include those of the hip, forearm, vertebral body, humerus, ribs and pelvis, while digital, skull, facial and fibular fractures are generally not osteoporotic in origin. The epidemiology of osteoporotic fractures reflects the importance of age and sex in the pathogenesis of the disease. The incidence of vertebral and hip fractures increases exponentially in later life in both men and women, although fracture rates at these sites in men increase about 10 years later (at age 75) than those in women.

Causes of osteoporosis

Box 31.2 sets out a list of risk factors for, and causes of, osteoporosis. Some of these (for example age and family history) are not modifiable. Whether or not an individual will develop osteoporosis depends on two factors: the peak bone mass, typically attained by age ~20 years, and the rate of loss of bone in later life. Probably the strongest influence on peak bone density is genetic – there is a

BOX 31.2 Risk factors for osteoporosis

Physiological

- Heredity: race, stature, family history
- Leanness
- Nutritional: very low calcium intake, hypovitaminosis D
- Inactivity
- Age

Pathological

- Endocrine: hypogonadism, hyperthyroidism, Cushing syndrome, primary hyperparathyroidism, adult growth hormone deficiency
- Gastrointestinal: partial gastrectomy, malabsorption
- Nutritional: prolonged undernutrition, e.g. anorexia nervosa
- Chronic liver disease
- Drugs: glucocorticoids, ciclosporin, thiazolidinediones, anticonvulsants, heparin, alcohol, tobacco
- Inflammatory disorders: rheumatoid arthritis, systemic lupus erythematosus, inflammatory bowel disease
- Haematological: myeloma, monocytic leukaemia, mastocytosis

close relationship between an individual's BMD and that of his parents, and bone density is reduced in the children of patients with osteoporosis.

Because of the evidence for a strong heritable component to peak BMD, there has been intense interest in identifying genes that regulate skeletal health. Thus, a large number of genome-wide association studies (GWAS) have been performed in recent years. These analyses have identified more than 20 genes that are associated with

alterations in BMD. Several of these genes cluster into four biological pathways: the vitamin D endocrine pathway, the oestrogen endocrine pathway, the wnt- β -catenin signalling pathway and the RANKL-RANK-OPG pathway. However, collectively these genes account for only a few percent of the variability in BMD. So, at present, it seems unlikely that genetic testing will improve the predictive power of algorithms used to identify patients at high fracture risk. Genetic factors may also underlie the significant racial differences in bone density and fracture incidence. Individuals of African or Polynesian ancestry have higher bone mass than Europeans and Asians.

Body weight is an important influence on bone density, with heavier people having greater bone mass, and lower fracture risk than those who are thin and/or underweight. The relationship between body weight and bone density is probably mediated in several ways, including mechanical loading of bone, the skeletal actions of hormones affected by nutritional status, and the actions of adipokines produced by adipose tissue.

Despite continued argument as to what constitutes an adequate calcium intake at different stages of life, the impact of calcium intake on bone density and fracture risk is probably slight at intakes >10 mmol (400 mg)/24 h (populations in Asia and Africa appear to maintain satisfactory bone health on intakes of <10 mmol/24 h). Vitamin D insufficiency (plasma 25-hydroxyvitamin D <50 nmol/L), or deficiency (<25 nmol/L), which is common in the contexts of skin pigmentation and sunlight deprivation, may contribute to reduced bone mass.

There has been extensive study of the relationship of physical activity to bone density. Cross-sectional studies are subject to various biases, but most prospective intervention studies indicate that this variable contributes only moderately to the differences in bone density that exist in postmenopausal women. However, prospective studies do indicate that weight-bearing exercise increases bone mass in the pre-pubertal skeleton.

Age has an important effect on bone density in both sexes, and also contributes to skeletal fragility, independently of BMD. In the proximal femur, loss of BMD begins in both sexes in the third and fourth decades, whereas in women BMD in the lumbar spine is probably stable until the perimenopause.

The loss of sex hormones has a deleterious effect on bone mass at whatever time it occurs. In both men and women it is oestradiol deficiency that appears to be critical. Men with inactivating mutations in the genes encoding either the aromatase or the oestrogen receptor have low BMD, despite the presence of normal or high testosterone concentrations. In women, late menarche and episodes of amenorrhoea with oestrogen deficiency reduce bone mass in much the same way as early menopause. Amenorrhoea with oestrogen deficiency can arise from a number of causes (Box 31.3). In the underweight, the low body weight itself has an effect on bone density. Adipose tissue is an important site of oestrone production, and high fat mass is associated with higher circulating concentrations of a number of bone anabolic factors including insulin, amylin and leptin. Cigarette smoking accelerates the metabolism of oestrogens to biologically inactive forms, is associated with reduced body weight

BOX 31.3

Causes of amenorrhoea and oestrogen deficiency in young women that are associated with reduced bone density

Ovarian failure

- Premature menopause
- Oophorectomy
- Chemotherapy

Hypothalamo-pituitary dysfunction

- Weight-related amenorrhoea
- Hyperprolactinaemia
- Drugs: long-acting gonadotrophin-releasing hormone agonists, depot medroxyprogesterone acetate contraception

and early menopause and may directly inhibit osteoblastic function: consequently, BMD tends to be lower in smokers than in non-smokers.

Various pathological conditions can also have significant effects on BMD (see Box 31.2). Alcohol appears to act as an osteoblast toxin and high intakes are also associated with both liver disease and hypogonadism. Glucocorticoids cause bone loss through inhibition of osteoblastic and osteocytic activity, intestinal calcium absorption and renal tubular calcium reabsorption. Glucocorticoids may directly or indirectly stimulate both hyperparathyroidism and bone resorption and, in men, are associated with reduced plasma testosterone concentrations.

Investigation and diagnosis

In recent years, there has been a move to characterize osteoporosis as a disease, principally so that it will be taken seriously by doctors, patients and regulatory authorities. However, this can be counterproductive to an understanding of the nature of the condition, since the process of bone loss is universal. Therefore, the key issue is often not so much one of diagnosis (required in the context of a disease), but one of assessing fracture risk, and then determining which interventions are cost-effective for a given level of risk. Fracture risk can be assessed from clinical risk factors (discussed above) and measurement of BMD.

Clinical risk factors for fracture

The risk factors of greatest importance include age, weight, family history and smoking. However, the most powerful predictor of future fractures is a past history of fracture. The presence of a deformed vertebra on a lateral spine or chest X-ray increases future fracture risk as much as five-fold, and a similar effect is seen with a past history of fractures at other sites. Therefore, treatment is sometimes indicated following fracture, even if the BMD is not particularly low.

Bone densitometry

There is a continuous, inverse, relationship between fracture risk and BMD. The best validated methodology for assessing BMD is dual-energy X-ray absorptiometry

(DEXA). Using this technology, a difference in bone density of one population standard deviation (about 10%) is associated with a two-fold difference in fracture risk. Most sites of bone density measurement give comparable prediction of global fracture risk. However, fracture at a given site is best predicted by bone density measurement at that site: for example, hip bone density predicts hip fracture risk better than measurements elsewhere. The sites most commonly assessed are the hip and spine. The fact that hip fractures are numerically the most important osteoporotic fractures argues in favour of using the hip as the preferred measurement site. However, the precision of measurement is greater in the spine, and the trabecular bone of the spine is more rapidly responsive to both disease and to therapeutic interventions. With advancing age, the value of vertebral scans diminishes because of artefacts associated with degenerative joint and disc disease. Because of these conflicting issues, it is common practice to measure bone density in both the lumbar spine and proximal femur. In the past, the forearm was a common site for measurement because that was all that was technically feasible. Modern densitometers can also measure the forearm, but there is no reason to prefer it to the spine and hip when these measurements are available.

Absolute values of bone density are not comparable between different anatomical sites, measurement techniques and make of densitometer. This problem has been addressed by reporting bone densities in terms of the young healthy population range, using standard deviation units (T-scores). The DEXA-based 'definition' of osteoporosis can be used as the threshold for intervention. However, it does not take into account the fact that clinical risk factors are multiplicative with the risk estimated from bone density measurement. For example, a T-score of -2.5 in an 80-year-old woman with a history of vertebral fracture is associated with a considerably higher risk of fracture over the next year than is the same density in a 50-year-old woman with no fracture history.

It is now standard practice to combine clinical risk factors with BMD in algorithms that produce estimates of absolute fracture risk. Examples, freely available via the internet, include the WHO-sponsored FRAX and the Garvan Fracture Risk Predictor. A 10-year hip fracture risk of $>3\%$ or an osteoporotic fracture risk of $>15\text{--}20\%$ are typically taken as indications for pharmaceutical intervention, but these thresholds are obviously influenced by the cost and safety of the interventions, balanced against their effectiveness. Of the other modalities available for assessment of BMD, computerized tomography (CT) scanning is rarely used because of the higher radiation exposure involved and ultrasound-based methods (e.g. of the os calcis) have not been as extensively validated.

Biochemical investigation

The investigations necessary in a patient who has osteoporosis will depend upon the extent to which the low bone mass can be accounted for by information already available. Thus, in an elderly subject or someone who has received long-term high-dose glucocorticoid therapy, there may be very little need for further investigations. However, in a younger individual whose bone density is

clearly subnormal without apparent cause, a more extensive investigation for an underlying abnormality is appropriate. The investigation thus depends on the clinical context (see Box 31.2). Measurements of sex hormone concentrations (particularly in men), thyroid, renal and liver function tests, plasma and urinary cortisol, plasma calcium, albumin and phosphate, 25-hydroxyvitamin D (25(OH)D), serum and urine protein electrophoresis, coeliac disease markers, serum tryptase (to diagnose mastocytosis) and full blood count may be appropriate.

There remains much debate about the utility of measuring bone turnover markers in the management of osteoporosis (Table 31.2). The potential uses include both prognosis (prediction of fracture, selection of patients for treatment, prediction of response to treatment) and therapeutics (selection of treatment, monitoring adherence to, and effectiveness of treatment). At present, there is little convincing evidence for use of biochemical markers in most clinical settings, but these are active areas of research and more data will emerge.

There has been concern that long-term suppression of bone turnover with potent bisphosphonates might increase the risk of osteonecrosis of the jaw and subtrochanteric femoral fractures. While the significance of these issues continues to be debated, 'drug holidays' after 5–10 years of bisphosphonate therapy are now commonly recommended; bone turnover markers have a role in deciding when treatment should be reintroduced.

There is limited value in measuring 24h urine calcium in patients with osteoporosis. The urine calcium excretion reflects an interaction of dietary calcium intake, fractional intestinal absorption, bone resorption and renal tubular reabsorption, and it is not possible to tease these apart without other investigations. The calcium:creatinine ratio in a fasting, second-voided urine sample in the morning is a poor guide to the rate of bone resorption, but does permit calculation of indices of renal calcium handling – though these are seldom needed in managing osteoporosis.

TABLE 31.2 Potential uses of bone turnover markers in osteoporosis

Clinical question	Utility of test
Diagnosis of osteoporosis or prediction of bone mass	None
Prediction of fracture and selection of patients for therapy	Possibly useful, but not yet validated
Selection of type of therapy	Theoretically appealing, not yet validated
Monitoring adherence to treatment	Useful
Monitoring effectiveness of therapy	Possibly useful, in conjunction with bone mass measurement
Predicting change in bone mass on treatment	Reasonable correlation in some studies
Determining duration of 'drug holiday' in patients on long-term antiresorptive therapy	Useful

In a few centres, the efficiency of intestinal calcium absorption is measured. However, there are practical difficulties in the performance of these tests (see Chapter 6) and there is no clear evidence that this approach is a useful guide to treatment selection or improves treatment outcomes.

Other investigations

The clinical history should include an assessment of recent and past dietary calcium intake. The presence of fractures should be confirmed radiologically and the radiographs scrutinized for evidence of other abnormalities, particularly the possibility of malignancy. Isotope scanning is also useful in detecting metastases. Bone biopsy is needed rarely, to diagnose pathologies such as osteomalacia or mastocytosis, and is reserved for patients with atypical presentations in whom there is a high index of suspicion of some other underlying abnormality (Appendix 31.1).

Treatment

Treatment of osteoporosis involves a number of strategies, all of which are aimed at reducing the individual's future risk of fracture. They can be divided into lifestyle changes (many of which are applicable to the whole population), oral calcium and vitamin D supplementation (in subjects with low intakes) and pharmacological measures. The latter are more expensive and more likely to be associated with side-effects, so are reserved for those with a moderately high absolute risk of fracture. In addition, minimization of exposure to drugs (particularly glucocorticoids) that are toxic to the skeleton is important.

Lifestyle modifications

Lifestyle interventions that can be recommended to all patients include maintenance of ideal body weight throughout life, maintenance of normal sex hormone concentrations from puberty until the late forties (as evidenced in women by regular menstruation), regular weight-bearing exercise, avoidance of smoking and avoidance of high alcohol intake. Exercise may be particularly important during childhood and adolescence, when the skeleton is very responsive to the loads placed upon it. Vigorous exercise interventions in postmenopausal women produce changes in bone density of only a few percent, but regular exercise, by increasing fitness, may also decrease an individual's risk of falling.

Calcium and vitamin D

While the fundamental building block of bone is collagen, calcium is also an important constituent, and its supply in the diet has been suggested to be a limiting factor for bone growth and maintenance. The average calcium intake in Western postmenopausal women is 17.5–22.5 mmol (700–900 mg)/24h, and there is evidence that increasing this by a further 25 mmol (1 g)/24h slows postmenopausal bone loss by ~50%, though calcium supplementation alone is not able completely to prevent

postmenopausal bone loss. Fracture risk appears to be reduced by ~10% with calcium supplements. However, the trials which suggest this fracture prevention also demonstrate an increase in risk of myocardial infarction of about 25% and of stroke of about 15%, effectively negating any potential benefit. These risks have not been demonstrated with *dietary* calcium intake, so modest augmentation of the diet may be a safer course to follow than the use of mineral supplements.

Vitamin D deficiency is extremely common among the frail elderly as a result of low levels of exposure to sunlight. If severe, it sometimes accelerates bone loss and it may also contribute to muscle weakness, increasing the risk of falls. If suspected, it can be confirmed by measurement of serum 25(OH)D. The current consensus is that optimum plasma concentrations of this compound are at least >50, and probably >70 nmol/L, although some authorities advocate concentrations as high as 100 nmol/L. Vitamin D deficiency can be prevented or treated with oral calciferol (at least 500–1000 U/day or, more conveniently, 50 000 U/month). There is evidence from randomized controlled trials in markedly deficient frail elderly subjects, that the provision of calcium and calciferol supplementation reduces hip fracture risk by more than a quarter. The use of the potent metabolites of vitamin D, such as calcitriol, is more expensive, more likely to cause hypercalcaemia or hypercalciuria and is of uncertain clinical benefit.

Pharmacological management

Bisphosphonates. The pharmacological management of osteoporosis is currently dominated by the bisphosphonates. These are relatively simple phosphate salts that have a very high affinity for the surface of bone but are very poorly absorbed from the gastrointestinal tract. Thus, only 1–2% of an oral dose is absorbed, about half of which is rapidly deposited on the bone surface, the balance being excreted unchanged in the urine. When osteoclasts resorb bone, they ingest the bisphosphonate and are effectively poisoned by it. This reduces the amount of bone resorption (bone resorption markers fall by 50–80%) with a consequent redressing of the imbalance between bone formation and resorption. Bisphosphonate remains present on the bone surface for many years and is gradually incorporated into the structure of the bone, so that it can inhibit remodelling cycles that occur years after the time of dosing. Thus, bisphosphonates have a long duration of action and intermittent administration is effective. Zoledronate, for instance, can be given as infrequently as once every 2–3 years by intravenous infusion and produces changes in bone density comparable to or greater than those seen after weekly oral dosing with alendronate.

The low oral bioavailability of bisphosphonates is a critical issue in their use. They must be taken fasting, with water alone, if they are to be absorbed at all. Aminobisphosphonates, such as alendronate and risedronate, can cause oesophagitis and gastric ulceration, so patients should not lie down for 30–60 min after oral dosing, since this may permit reflux of the tablet into the oesophagus. This does not appear to be such a problem for the less potent, non-amino bisphosphonate, etidronate.

A common phenomenon after first exposure to intravenous (and occasionally oral) nitrogen-containing bisphosphonates is an 'acute phase response' of fever, myalgia and headache that lasts 24–48 h. This is mediated by the release of TNF α and interleukin (IL)-6 from peripheral blood T cells. While sometimes unpleasant, the reaction is transient, and tends not to recur with subsequent doses.

Potent bisphosphonate use in postmenopausal women results in an increase in bone density comparable to that produced by hormone replacement therapy. Bone resorption is reduced by more than one half, and the risk of hip, spine and forearm fractures is lowered by ~50%. While there are a number of trials assessing the effects of bisphosphonates on fracture risk over periods of three to four years, there are only limited data available from longer-term studies. The available data suggest that anti-fracture efficacy is maintained with long-term use.

Of the currently available bisphosphonates, the evidence for anti-fracture efficacy is strongest for alendronate (most conveniently given as a weekly dose), risedronate (weekly) and intravenous zoledronate (annually), each of which has been demonstrated to prevent both vertebral and non-vertebral fractures in osteoporotic and glucocorticoid-treated patients and in men as well as postmenopausal women. Thus, these bisphosphonates are currently the first choice therapy in all forms of osteoporosis.

Following discontinuation of bisphosphonates, bone resorption rises to some extent, but even several years after discontinuation of long-term use of alendronate and zoledronate, bone turnover markers do not return to baseline. The same is true of bone density, which declines only slowly after withdrawal of these agents. This prolonged duration of action is a consequence of the long residence time of bisphosphonates in bone, and it raises important (as yet unanswered) questions about the optimal duration of continuous therapy. There is general agreement that these agents should be used continuously for periods of up to five years. Beyond that, some experts reduce the dose, whereas others opt either for no change or for a period off treatment. There is no clear evidence at present to determine what is optimal. Risedronate and ibandronate have a more rapid offset of effect, and will often require re-institution of therapy within 6–12 months of treatment cessation.

Hormone replacement therapy. Hormone replacement therapy (HRT) was the most widely used intervention for the prevention and treatment of osteoporosis for many years, but was displaced by the aminobisphosphonates, because of their demonstrated anti-fracture efficacy, and because the Women's Health Initiative produced a marked change in attitudes to hormone replacement therapy. On the one hand, it clearly demonstrated that HRT reduces the incidence of fractures, including hip fractures. On the other, it raised concerns regarding safety, particularly with respect to cardiovascular and cerebrovascular disease, breast cancer and venous thromboembolism. Public and professional enthusiasm for the use of HRT in the management of osteoporosis waned, because the older age group in whom osteoporotic fracture risk is high is the same group in whom vascular disease is a major concern.

If early postmenopausal women require interventions for osteoporosis, HRT may still be a reasonable option, although the balance of risks and benefits of HRT in younger (<50 years) postmenopausal women is uncertain, and the bisphosphonates are also suitable for use in this context. The cardiovascular and skeletal effects of transdermal oestrogen preparations have not been assessed in large randomized studies. Unlike the bisphosphonates, oestrogen effects on bone disappear within months of treatment discontinuation, so long-term reductions in fracture risk are unlikely to occur.

Selective oestrogen receptor modulators. Selective oestrogen receptor modulators (SERMs) are a newer class of pharmaceuticals with mixed oestrogen agonist/antagonist activities that vary from tissue to tissue. Thus, the prototypic SERM, raloxifene, acts as an oestrogen agonist on bone, but as an antagonist in the breast and endometrium. Raloxifene reduces bone resorption and increases bone density, but is less potent than either oestrogen or bisphosphonates in this regard. It decreases the incidence of vertebral fractures but it has not been shown to have any effect on non-vertebral fractures. It substantially reduces the incidence of breast cancer, and does not significantly influence risk of vascular disease. Basodoxifene has similar effects and is being developed for co-administration with oestrogen, thus augmenting its bone effects, while the anti-oestrogen actions of the SERM block the proliferative effects of the oestrogen on breast and endometrial cells. Lasofoxifene is the only SERM to be shown to reduce risk of non-vertebral factors, but it is not clear that this drug will reach the market.

Parathyroid hormone. Parathyroid hormone (PTH) preparations, either PTH(1-34) (teriparatide) or PTH(1-84), were the first anabolic therapies to be approved for treatment of osteoporosis. When given as a daily injection, they produce substantial increases in BMD at skeletal sites containing predominantly trabecular bone, with less marked increases in cortical BMD, and reductions in fracture incidence that tend to be greater than are seen with other currently available agents. PTH use is restricted to a period of 1.5–2 years because of the high incidence of osteosarcoma in rat toxicology studies, although this problem has not been an issue in human use. Unlike the antiresorptive agents, it causes sustained *increases* in markers of both formation and resorption. Practices for combining PTH with antiresorptives vary; co-administration with alendronate may blunt its anabolic effects, though this appears not to be the case for combination with zoledronate. The anabolic effects dissipate rapidly after treatment discontinuation, so locking in gains in BMD with subsequent antiresorptive therapy is important.

Other pharmacological treatments. The divalent cation strontium (in the form of strontium ranelate) has been shown to prevent both vertebral and non-vertebral fractures, though effects on hip fracture were only seen in *post hoc* analyses. The mechanism of action of strontium on bone is not understood, but some of the increase in

BMD it produces is artefactual, resulting from replacement of calcium ions by the heavier strontium ions in the hydroxyapatite. BMD measurements in strontium-treated patients, should be adjusted for skeletal strontium content.

Denosumab is a monoclonal antibody against the osteoclastogenic factor, RANKL. It produces profound reductions in both bone resorption and bone formation, resulting in increases in BMD and decreases in fracture rates comparable to those resulting from the use of potent bisphosphonates. It is given by six-monthly subcutaneous injection but there is a rapid loss of effect 6–8 months after an injection, with an overshoot in bone turnover during this time. Therefore, maintaining continuity of treatment is very important to achieve anti-fracture efficacy.

Cathepsin K is the principal osteoclastic enzyme involved in the degradation of type I collagen. It is therefore an obvious target for an antiresorptive agent. Several early studies in this area were abandoned because of insufficient specificity of the investigation compounds. The cathepsin K inhibitor odanacatib, has demonstrated sustained year-on-year increases in bone density in phase 2 studies. A large phase 3 study to determine anti-fracture efficacy is nearing completion. Tartrate resistant 5b concentrations are not suppressed (as osteoclasts remain numerous on the bone surface, although their bone resorbing activity is markedly curtailed) and the suppression of bone turnover markers with odanacatib is less marked than seen with other antiresorptive agents; in particular the suppression of bone formation appears to diminish after 2–3 years. Discontinuation of treatment is accompanied by a rapid rise in bone turnover markers.

In recent years, the wnt-LRP5 pathway has emerged as a critical regulator of bone formation. Activity of this pathway is regulated by sclerostin, an inhibitor secreted by the osteocyte. A number of studies are exploring the efficacy of monoclonal antibodies directed against sclerostin as bone anabolics. Phase 1 and 2 studies suggest that these could be the most effective bone anabolic drugs yet developed; results of phase 3 trials are awaited.

In the past, sodium fluoride and calcitonin were widely used to treat osteoporosis but good evidence of anti-fracture efficacy is lacking for both and fluoride interferes with normal bone mineralization. As a result, both agents have dropped out of common usage.

Biochemical responses to treatments

Antiresorptive agents such as HRT, SERMs and the bisphosphonates produce an early (within weeks) reduction in indices of bone resorption, followed by a slower fall in indices of bone formation (Fig. 31.7). Similar effects are observed in hypogonadal men treated with testosterone replacement. These effects are sustained for the duration of therapy and, in the case of potent bisphosphonates, which have long skeletal retention times, may persist for considerable periods of time (months or years) after therapy is stopped. The most potent currently available bisphosphonates may reduce markers of bone resorption by up to 80% from baseline levels. Denosumab can produce even larger decreases in bone resorption, most patients having undetectable resorption markers in

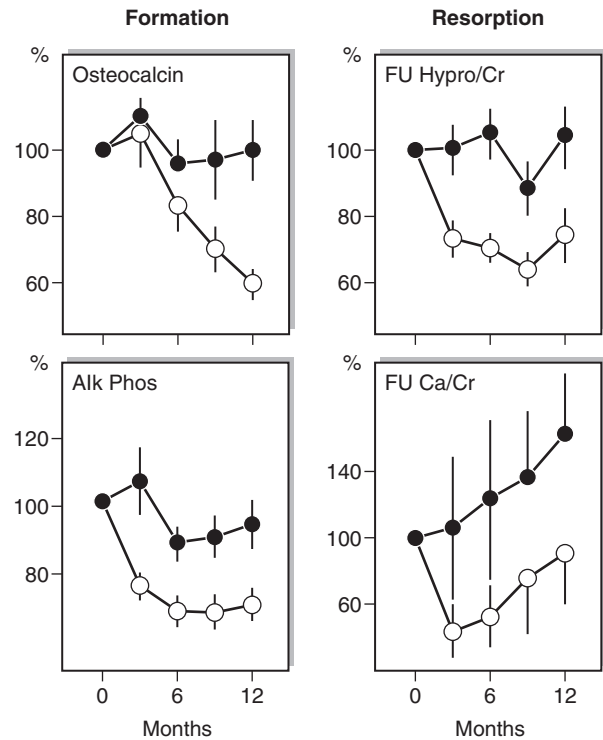


FIGURE 31.7 ■ Changes in mean (\pm SD) biochemical indices of bone formation and resorption in postmenopausal women receiving either placebo (closed circle) or a regimen of continuous oestrogen and progesterone (open circle). The group receiving active treatment show significant suppression of both bone resorption and formation. There is a tendency to reach a plateau in these indices as a new steady state is reached. Note that the variance in fasting urine calcium/creatinine ratios (FU Ca/Cr) is considerably greater than that in fasting urine hydroxyproline/creatinine (FU Hypro/Cr). (Data from Christiansen C, Riis BJ 1990 17 β estradiol and continual norethisterone: a unique treatment for established osteoporosis in elderly women. *Journal of Clinical Endocrinology & Metabolism* 71:836–841.)

the first weeks after injection. In contrast, the cathepsin K inhibitor odanacatib only reduces resorption markers by about 50%, with smaller declines in formation markers. Plasma calcium and phosphate concentrations can also fall during the early phase of antiresorptive therapy but usually recover without specific intervention.

Intermittent PTH treatment produces biochemical changes characterized by an initial increase in markers of bone formation (P1NP, e.g. typically doubling), reflecting primary activation of osteoblasts. A smaller increase in osteoclastic activity follows as a secondary, coupled phenomenon. Plasma 1,25(OH)₂D concentration increases by ~25% and there may be small falls in 25(OH)D. Plasma calcium concentration may increase and plasma phosphate decrease during treatment because PTH also stimulates renal tubular calcium reabsorption and phosphaturia (directly) and intestinal calcium absorption (indirectly). Occasionally, a reduction in PTH dose may be required because of mild hypercalcaemia. Early data with anti-sclerostin antibodies demonstrate large early increases in formation markers and, surprisingly, some inhibition of bone resorption markers. Over a period of months, both changes tend to revert towards baseline values.

Treatment with strontium produces inconsistent changes in markers, smaller in magnitude than most agents discussed above. Notably, strontium interferes in many routinely available assays for calcium concentration.

Growth hormone replacement in adults who are deficient improves bone mass, primarily by stimulating cortical bone accretion. Growth hormone treatment induces an early increase in bone resorption markers, and later increases (~two-fold) in bone formation markers that reach a peak after a year of treatment, and then gradually decline. The peak in osteocalcin is delayed relative to the other formation markers.

OSTEOMALACIA

Osteomalacia is the term used to describe the disorder arising from defective mineralization of bone. It is a histological diagnosis, based on the volume of unmineralized osteoid, its extent over bone surfaces, its thickness and the rate of mineralization, as assessed by tetracycline labelling. If a mineralizing defect is a diagnostic possibility in an unusual case, then biopsy should be undertaken (Appendix 31.1). In adults, the radiographic appearances are generally non-specific, showing generalized osteopenia or (in severe cases) vertebral collapse. Looser zones (short radiolucent lines through the cortex, at right angles to the shaft) are very suggestive of osteomalacia, but occur in only a minority of patients.

When osteomalacia arises before growth is complete, the clinical and radiographic features differ from those seen in adulthood, and the condition is termed rickets. The histological processes occurring in osteomalacia and rickets are similar. In order for newly formed osteoid to mineralize, adequate supplies of mineral (calcium and phosphate) must be available and the function of the osteoblasts that regulate the mineralization process must be normal. The causes of osteomalacia are conveniently categorized into those where there is inadequate supply of mineral and those where osteoblast function is defective. The former group may be further categorized into those where inadequate calcium supply is the critical factor (calciopenic) and those where inadequate phosphate supply is critical (phosphopenic). Calciopenic forms are frequently associated with abnormalities of vitamin D metabolism. Biochemical investigations can suggest that osteomalacia is present and indicate possible aetiologies, but cannot prove its presence. However, the biochemical findings in calciopenic and phosphopenic forms of osteomalacia are often very characteristic, so that bone biopsy is not necessary in routine cases.

Calciopenic osteomalacia

The major causes of calciopenic osteomalacia are summarized in Box 31.4.

Vitamin D deficiency

Rickets and osteomalacia were endemic in the UK and many other European countries from the 17th century until the mid-20th century. Crucial scientific observations

BOX 31.4 Causes of calciopenic osteomalacia

Low 25(OH)D, normal 1,25(OH)₂D synthesis

- Low calcium intake
- Binding of dietary calcium by phytate and other factors
- Lack of exposure to sunlight
- Post-gastrectomy
- Malabsorption syndromes (small intestinal disease, hepatobiliary disease, pancreatic disease)
- Anticonvulsants
- Mutation in gene for 25-hydroxylase (*CYP2R1*)

Normal 25(OH)D, reduced 1,25(OH)₂D synthesis

- Chronic kidney disease
- Vitamin D-dependent rickets type I

Normal 25(OH)D, normal 1,25(OH)₂D synthesis

- Vitamin D-dependent rickets type II

made in post-war Vienna in 1919–1922 established the role of vitamin D in the aetiology of privational osteomalacia. Although it was clearly demonstrated at that time that it was possible to cure privational rickets by exposing children to sunlight, as well as by the administration of a fat-soluble food factor, the dominance of nutritional theory led to the active factor being termed a vitamin, that is, an obligatory food component that the body cannot synthesize. It is now clear that under the influence of ultraviolet light, the skin can synthesize vitamin D₃ (cholecalciferol), the precursor of the prohormone 25(OH)D, and that dietary sources of vitamin D (mainly vitamin D₂, ergocalciferol) are critical only when exposure to sunlight is significantly limited.

Diminished exposure to ultraviolet light through industrial pollution and the practice of infant swaddling may have been important during the years that rickets was endemic, and would account for its greater prevalence in more northerly cities in the UK. In the present day, vitamin D-deficient osteomalacia is observed most commonly in individuals who are sunlight deprived because of skin pigmentation, whose cultural practices involve sunlight deprivation, the malnourished and the institutionalized elderly. Osteomalacia occurring in patients with malabsorption syndromes or after partial gastrectomy is not so easily explained, but does suggest that a nutritional factor is important. This factor is most probably calcium.

Reduced intake of calcium, or impaired calcium absorption, initially causes a small fall in plasma ionized calcium concentration and compensatory secondary hyperparathyroidism. As a consequence of this, a greater proportion of 25(OH)D is metabolized to 1,25(OH)₂D – the active hormone – rather than to inactive metabolites such as 24,25(OH)₂D₃ and 25,26(OH)₂D₃. In the short term, the increased 1,25(OH)₂D production, by increasing fractional absorption of calcium from the intestine, corrects the hypocalcaemia at the expense of modest hyperparathyroidism. However, 1,25(OH)₂D has an additional effect upon the hepatic metabolism of 25(OH)D, accelerating the metabolic clearance of 25(OH)D to its inactive dihydroxylated forms (see Fig. 6.3, p. 97). In the chronic situation, the sustained increase in 1,25(OH)₂D

production eventually depletes the available 25(OH)D pool. Once the concentrations of 25(OH)D have fallen beyond a critical point, 1,25(OH)₂D synthesis itself may become insufficient to sustain adequate calcium absorption, so that there is overt hypocalcaemia. Plasma phosphate is also reduced because of diminished intestinal absorption and the renal effects of hyperparathyroidism. With insufficient calcium and/or phosphate available for mineralization, osteomalacia follows.

Osteomalacia and low plasma 25(OH)D concentrations are also found in some patients on long-term anticonvulsant therapy. Hepatic enzyme induction by the anticonvulsant drugs and accelerated clearance of 25(OH)D are important, but the people most likely to develop this syndrome are institutionalized patients who also have limited exposure to sunlight. Intestinal calcium transport can be impaired directly by phenytoin. A patient from Nigeria with severe rickets/osteomalacia and a very low serum 25(OH)D concentration was found to be homozygous for an inactivating mutation in the gene *CYP2R1*, that encodes the enzyme mainly responsible for 25-hydroxylase activity.

Osteomalacia in people of South Asian origin is multifactorial. As well as diminished dermal synthesis of vitamin D, the traditional diet has a high content of calcium-binding factors such as phytate, which limit the bioavailability of dietary calcium. In groups with a high prevalence of osteomalacia, the disease may be exacerbated by lack of sunshine in winter and at times of increased vitamin D requirements: infancy, adolescence and pregnancy.

The clinical features of privational osteomalacia in adults include bone pain, proximal myopathy and fractures. Bone pain and myopathy are often overlooked and their severity only appreciated when a cure has been effected. The aetiology of the myopathy is unknown. It is not a feature of hypocalcaemia, and does not invariably occur in other types of osteomalacia. In children, rickets may additionally result in difficulty with walking, growth retardation, bowed legs and enlargement of the costochondral junctions.

Defects in 1,25-dihydroxyvitamin D synthesis or action

Osteomalacia and rickets can result from deficiency of the renal 1 α -hydroxylase enzyme and thus failure to synthesize 1,25(OH)₂D, or from defective action of 1,25(OH)₂D, owing to abnormalities of the receptor. Deficiency of the renal 1 α -hydroxylase enzyme can be acquired, as in chronic kidney disease (CKD, see below), or can occur as a rare autosomal recessive disorder known as vitamin D-dependent rickets (VDDR) type I. Clinically, these children develop florid rickets in the first two years of life. The syndrome of VDDR type II is also a rare autosomal recessive disorder, often associated with alopecia totalis, but otherwise with a clinical picture similar to VDDR type I. It is caused by an abnormality in the receptor for 1,25(OH)₂D, arising from mutations either in the DNA-binding or the hormone-binding domain, and there is target organ resistance to the hormone. Acquired resistance to 1,25(OH)₂D has been seen in untreated coeliac disease, where the intestinal mucosa is very abnormal and the absorption of calcium and phosphate impaired.

Laboratory investigation

The relevant laboratory investigations are summarized in Table 31.3. It is important to note that the plasma calcium concentration may be normal in the early phase of privational vitamin D deficiency. Hypophosphataemia due to hyperparathyroidism and reduced intestinal phosphate absorption can occur in any type of calciopenic osteomalacia. However, there are occasional patients with an apparent renal resistance to the effects of PTH who do not become hypophosphataemic when deficient in 25(OH)D. These patients have marked hypocalcaemia and hyperphosphataemia with increased plasma ALP activities.

All bone turnover markers, apart from plasma osteocalcin, are increased in vitamin D deficiency osteomalacia. When assessing plasma phosphate and ALP concentrations, it is important to bear in mind the normal age-related ranges for these indices. The key diagnostic investigation in privational

TABLE 31.3 Typical biochemical and histological findings in calciopenic and phosphopenic forms of osteomalacia and rickets

	Plasma calcium	Plasma phosphate	Urinary TmP/GFR	Plasma PTH	Plasma alkaline phosphatase	Plasma 25(OH)D	Plasma 1,25(OH) ₂ D	Bone histology
Vitamin D deficiency:								
early	N	↓/N	↓	↑	↑	↓	↑	HPT
late	↓	↓/N	↓	↑↑	↑↑	↓↓	↓/N	OM, HPT
Vitamin D-dependent rickets type I	↓	↓/N	↓	↑↑	↑↑	N	↓↓	OM, HPT
Vitamin D-dependent rickets type II	↓	↓/N	↓	↑↑	↑↑	N	↑↑	OM, HPT
Renal phosphate leak (high FGF23)	N	↓↓	↓↓	N	↑	N	↓	OM
Renal phosphate leak (low FGF23)	N	↓↓	↓↓	N	↑	N	↑	OM
Antacid-induced hypophosphataemia	↑/N	↓↓	↑	↓	↑	N	↑	OM

FGF, fibroblast growth factor; HPT, hyperparathyroidism; N, normal; OM, osteomalacia.

vitamin D deficiency is measurement of serum 25(OH)D concentration. Measurement of 1,25(OH)₂D is indicated in infants with rickets who have no medical or social risk factors for privational vitamin D deficiency and who are hypocalcaemic despite normal plasma 25(OH)D concentrations.

Responses to therapy

Vitamin D deficiency should be treated with calciferol. Single high-dose treatment (300 000–600 000 units orally in an adult) may be more effective than low-dose long-term replacement and reduces problems of adherence to therapy. Patients with chronic conditions can be given weekly or monthly treatment with smaller doses, or annual prophylaxis in mid-autumn. There is a marked tendency for spontaneous improvement in the spring and summer. Attention needs also to be directed to underlying problems such as malabsorption. Vitamin D therapy is ineffective if the dietary calcium content is very low.

Effective therapy is followed by changes in the plasma biochemistry. Plasma calcium concentration rises and hyperparathyroidism is corrected. A marked rise in plasma phosphate concentration occurs after initiation of therapy and peaks at 3–5 weeks; thereafter plasma phosphate gradually declines, reaching the concentration which is normal for that individual after about 6–9 months. Plasma 1,25(OH)₂D concentration is also supranormal during this period, but there is no clinical need to measure 1,25(OH)₂D in this circumstance. Plasma ALP activity often flares during the first two to three weeks but then declines exponentially with a half-time of 50–100 days. A similar early increase in plasma concentrations of procollagen 1 extension peptides can also be seen. Osteocalcin concentrations also rise transiently, and peak about two weeks after treatment is started.

Vitamin D-dependent rickets type I is best treated with oral calcitriol (50 ng/kg per day) or alfacalcidol (100 ng/kg per day). Vitamin D-dependent rickets type II may respond to a combination of calcium supplementation and treatment with very high doses of calcitriol (60 µg/kg per day) or alfacalcidol (100 µg/kg per day).

Any patient being treated continuously with *pharmacological* doses of any form of vitamin D is at risk of vitamin D toxicity, and plasma calcium concentrations need to be monitored and dosages adjusted accordingly. Toxicity is particularly liable to occur as bone disease heals and therapeutic requirements fall.

Phosphopenic osteomalacia

In this group of conditions, the low plasma phosphate concentration is the main factor underlying the osteomalacia. Very rarely, this may result from the ingestion of phosphate-binding antacids, but most commonly it is due to renal tubular dysfunction and phosphaturia (manifest as a low TmP/GFR, see Table 31.3). The disorders can be classified into genetic and acquired forms and by whether or not there are other associated renal tubular defects (aminoaciduria, bicarbonate loss with systemic acidosis, glycosuria and hypokalaemia: the Fanconi syndrome). The most frequently encountered forms are listed in Box 31.5, and their aetiology discussed in Chapter 6.

BOX 31.5 Osteomalacia and rickets resulting from renal tubular phosphate loss

Isolated phosphate leak

Hereditary

- X-linked hypophosphataemic rickets – X-linked dominant (307800 – *PHEX*)
- Autosomal dominant hypophosphataemic rickets (193100 – *FGF23*)
- Hypophosphataemic rickets with hypercalciuria – autosomal recessive (241530 – *SLC34A3*)
- X-linked hypophosphataemic rickets (Dent disease) – X-linked recessive (182309 – *CLCN5*)
- Hypophosphataemic rickets – autosomal recessive (241520 – *DMP1*)

Acquired

- Tumour-associated: mesenchymal cell tumours, polyostotic
- Fibrous dysplasia, neurofibromatosis
- Drug-associated: imatinib, ifosfamide, saccharated ferric oxide

Generalized tubular disorders (Fanconi syndrome)

Hereditary – many different types with dominant, recessive or X-linked inheritance, such as:

- Cystinosis (219800 – *CTNS*)
- Wilson disease (277900 – *ATP7B*)
- Tyrosinaemia, type I (276700 – *FAH*)
- Oculocerebrorenal syndrome/Dent disease-2 (300555 – *OCRL*)
- Hereditary fructose intolerance (229600 – *ALDOB*)
- Fanconi–Bickel syndrome (227810 – *GLUT2*)

Acquired

- Renal tubular acidosis (type 1 and type 2)
- Heavy metal toxicity: lead, cadmium, mercury, platinum, uranium
- Drugs: out-dated tetracycline, cisplatin, 6-mercaptopurine, amphotericin, trimethoprim, pentamidine, toluene
- Dysproteinaemias: myeloma, light chain disease, amyloidosis
- Autoimmune-mediated: systemic lupus erythematosus, Sjögren syndrome, rheumatoid arthritis
- Interstitial nephritis, renal transplantation

Numbers in brackets after genetic disorder indicate numbering on Online Mendelian Inheritance in Man (OMIM); the gene for each disorder is in italics.

The presentation of the inherited syndromes is usually in infancy with a clinical picture similar to that of rickets due to privational vitamin D deficiency. The most common form of inherited hypophosphataemic rickets, the X-linked type, is exceptional in that myopathy is absent and that marked ligamentous calcification can occur. Male children are more severely affected than females and short stature is a prominent feature.

Adult onset hypophosphataemic osteomalacia is usually associated with mesenchymal tumours (oncogenic osteomalacia) that are often difficult to identify because of their small size. These tumours secrete phosphatonins, such as FGF23, that act on the renal tubules to induce phosphaturia. If a tumour is solitary and can be removed

surgically, then complete resolution of the osteomalacia occurs. Tumours can be identified by either magnetic resonance imaging (MRI) or scintiscanning using isotope-labelled somatostatin or sestamibi, but in a significant minority of cases even these sophisticated techniques may be inadequate. Plasma concentrations of FGF23 are usually increased, and fall after successful resection of the tumour.

Heavy metal nephropathy or light chain deposition in the renal tubules can cause an acquired form of the Fanconi syndrome with hypophosphataemic osteomalacia.

Laboratory investigation

The occurrence of rickets or osteomalacia in an individual without evidence of vitamin D deficiency should prompt consideration of phosphopenia as a cause. A key, and often overlooked, investigation is measurement of the plasma phosphate which, if low, should prompt evaluation of renal phosphate handling. The typical laboratory findings are listed in Table 31.3. The critical biochemical features that distinguish osteomalacia caused by hypophosphataemia due to a renal phosphate leak from that caused by vitamin D deficiency are:

- normal plasma 25(OH)D concentration
- low TmP/GFR, which persists after treatment of the bone disease
- absence of hyperparathyroidism.

Plasma FGF23 measurements are helpful in the diagnosis of oncogenic osteomalacia and some of the genetic disorders. Plasma calcitriol concentrations are usually low in hypophosphataemic osteomalacia that is the result of excess FGF23 activity, but high in conditions such as Dent disease or hereditary hypophosphataemic rickets with hypercalciuria, where the primary defect lies in renal tubular phosphate transporters.

Special tests are required for the diagnosis of the rare inherited and acquired metabolic disorders in which hypophosphataemic osteomalacia is part of a more generalized Fanconi syndrome.

Treatment

When hypophosphataemic osteomalacia occurs with inappropriately low concentrations of calcitriol (usually in the context of increased FGF23 activity) then it is best treated with phosphate supplements and 1 α -hydroxylated vitamin D derivatives. Commercial preparations of sodium dihydrogen phosphate, are available but the precise form of phosphate is not important. Phosphate (total 30–120 mmol/day) has to be given frequently through the day to be effective, because of its rapid loss through the kidneys. Phosphate therapy alone induces hyperparathyroidism, so vitamin D therapy, which offsets this and enhances intestinal phosphate absorption, is also used. These patients are vitamin D resistant, so if calciferol (vitamin D₂) is used, pharmacological doses are required and there is a danger of prolonged vitamin D toxicity. Alfacalcidol or calcitriol are therefore preferred to the parent compound. The histological responses to alfacalcidol or calcitriol also seem to be superior to those seen with calciferol. The therapy should aim to keep plasma

phosphate around 1.0 mmol/L (measured 1–2 h after a dose), but to avoid hypercalciuria and hypercalcaemia. There is an increased risk of metastatic calcification, particularly in the kidney. The risk of nephrocalcinosis is related primarily to the amount of phosphate given, and its occurrence suggests that the dose of phosphate should be reduced (or stopped in the X-linked form when growth is complete). Healing of the bone disease is accompanied by a fall in plasma ALP activity.

Hereditary hypophosphataemic rickets with hypercalciuria is caused by mutations in the renal tubular phosphate transporter NaPi2c. FGF23 concentrations are not increased and plasma calcitriol is appropriately raised (given the hypophosphataemia). The bone disease can be cured with phosphate alone. In some conditions, the Fanconi syndrome responds to treatment of the underlying disorder (e.g. Wilson disease).

Antacid-induced hypophosphataemic osteomalacia is treated by withdrawal of antacids and the short-term prescription of phosphate supplementation.

Osteomalacia and acidosis

Longstanding metabolic acidosis is also associated with osteomalacia. The clinical situations in which this occurs most commonly are types 1 and 2 renal tubular acidosis and patients with bladder resection and ureterosigmoidostomy. The physiological basis of the osteomalacia is incompletely understood. At high [H⁺], the renal 1 α -hydroxylase enzyme is inhibited, so the increase in plasma 1,25(OH)₂D concentrations induced by phosphate depletion is less than expected. However, metabolic acidosis reduces TmP/GFR and, in the short term, stimulates 1,25(OH)₂D synthesis. The mechanism of induction of osteomalacia may thus be similar to that seen in simple vitamin D deficiency. In keeping with this, plasma 25(OH)D concentrations have been reported to be low in many of the published cases of osteomalacia associated with ureterosigmoidostomy. In type 2 renal tubular acidosis, the hypocalcaemia and hypophosphataemia accompanying the Fanconi syndrome could also be important contributory factors. Treatment with alkali alone (sodium bicarbonate) can improve the osteomalacia, but in the limited studies published, has not produced complete healing.

Defective osteoblast function and osteomalacia

Osteomalacia can arise from either congenital or acquired abnormalities of osteoblast function. The acquired forms relate to aluminium or the use of high-dose etidronate in the treatment of Paget disease. In both these conditions, osteoblast numbers are reduced, plasma ALP activity is low and the osteomalacia is resistant to treatment with vitamin D.

Fluoride therapy, or endemic fluorosis, produces a mineralization defect despite abundant osteoblasts being present.

Hypophosphatasia. This is a rare familial disorder with autosomal recessive inheritance arising from mutations in the tissue non-specific ALP gene (*ALPL*). There are varying degrees of clinical severity, depending on the

TABLE 31.4 Clinical syndromes of hypophosphatasia

Presentation	Clinical features
Perinatal	Stillbirth or neonatal death; absence of mineralized skeleton
Infantile	Diagnosed <6 months; rickets, flail chest, hypercalcaemia/hypercalciuria, pyridoxine-responsive seizures
Childhood	Premature loss of primary teeth, rickets, hypercalcaemia/hypercalciuria
Adult	Middle age; recurrent stress fractures; pathological fractures
Odontohypophosphatasia	Premature loss of primary teeth only

types of mutation and the residual activity of the enzyme. The clinical presentation can vary from a lethal perinatal form through to a relatively benign disorder presenting in adulthood (Table 31.4). Heterozygous carriers can be asymptomatic or have a mild form of the disease. In the childhood form, the diagnosis is made usually after the age of six months of age because of orthopaedic abnormalities that resemble rickets. Other features include premature loss of the deciduous teeth, premature closure of the cranial sutures and growth retardation.

The characteristic biochemical findings are reduced plasma ALP activity (both total and bone ALP) with concomitant increases in endogenous substrates of the enzyme. These include plasma pyridoxal 5'-phosphate, inorganic pyrophosphate and phosphoethanolamine. An elevated plasma pyridoxal 5'-phosphate concentration is specific to hypophosphatasia (provided the patient is not taking vitamin B supplements), but urine phosphoethanolamine excretion can be modestly increased in a variety of other disorders. Severely affected children may have hypercalcaemia and hypercalciuria. Plasma phosphate concentrations also tend to be above average for age (due to an increase in the TmP/GFR), and PTH concentrations relatively low. Antenatal diagnosis of the severe form is possible by measurement of amniotic fluid ALP.

The bone shows osteomalacia with reduced osteoblast numbers. Treatment options are limited, but there are case reports of improvement of bone disease with parathyroid hormone treatment, and in severe infantile disease with bone marrow transplantation. The most promising treatment for severe hypophosphatasia is enzyme-replacement therapy, using a fusion protein (named *asfotase alfa*) that comprises the TNSALP ectodomain, the constant region of the human IgG1 Fc domain and a deca-aspartate motif.

CHRONIC KIDNEY DISEASE – MINERAL AND BONE DISORDER

Bone diseases complicating CKD emerged as major problems with the development of dialysis programmes, which permitted the prolonged survival of patients with

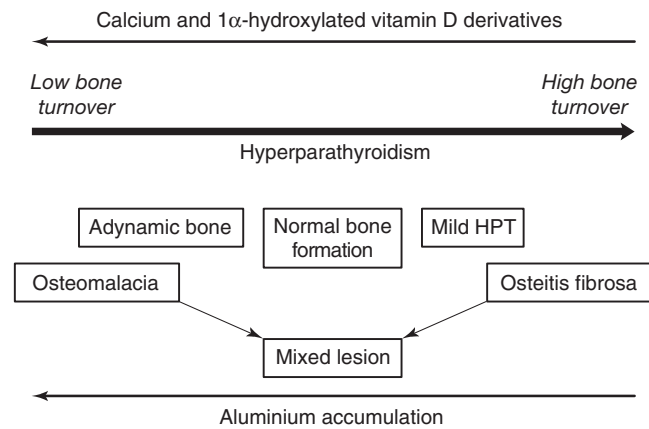


FIGURE 31.8 ■ Chronic kidney disease mineral and bone disorder is a dynamic process. The main influences determining the nature of the bone histology are the degree of hyperparathyroidism (HPT) and the degree of aluminium accumulation in bone. Hyperparathyroidism can be suppressed to varying degrees by the manipulation of dietary and dialysate calcium and the use of 1 α -hydroxylated vitamin D derivatives, or by parathyroidectomy. Aluminium accumulation also favours low bone turnover and, in significant quantities, can cause osteomalacia.

established renal failure. The major aetiological factors are disturbances of the parathyroid hormone–calcitriol axis and skeletal retention of aluminium. Renal bone disease is not uniform in its features because the effects and interactions of these factors differ at various stages of its evolution. The disease can be modified substantially by a variety of treatments, from changes in dialysate composition, diet or drugs to renal transplantation. Chronic kidney disease – mineral and bone disorder (CKD-MBD), previously called renal osteodystrophy, can be regarded as a dynamic process that is affected by both patient characteristics and a number of pathophysiological and iatrogenic factors. These are illustrated in Figure 31.8.

Aetiology

Parathyroid hormone–calcitriol–FGF23 axis

Mild hyperparathyroidism can be detectable when the glomerular filtration rate falls to <50 mL/min. The hyperparathyroidism is secondary to a modest fall in plasma ionized calcium, resulting from a reduction in intestinal calcium absorption that, in turn, is due to a decline in the production of 1,25(OH)₂D because of reduced renal 1 α -hydroxylase activity. This arises initially not because of the loss of functioning nephrons, but because of phosphate retention. Plasma phosphate concentrations are not increased and may even be low because of secondary hyperparathyroidism, but there is saturation of the intracellular phosphate pool that ‘senses’ phosphate requirements. The high intracellular phosphate is responsible for the reduction in renal 1 α -hydroxylase activity and therefore 1,25(OH)₂D synthesis. Rising FGF23 concentrations in response to phosphate retention may mediate this early reduction in 1 α -hydroxylase activity: increases in plasma FGF23 concentrations seem to precede the development of hyperparathyroidism. Reduction in the intake or the intestinal absorption of phosphate at this early

stage reverses all these abnormalities: $1,25(\text{OH})_2\text{D}$ concentration in the blood rises, intestinal calcium absorption increases and hyperparathyroidism is suppressed.

As kidney disease progresses, the disturbances in the PTH- $1,25(\text{OH})_2\text{D}$ axis become more marked. When the glomerular filtration rate falls to <30 mL/min, hyperphosphataemia develops, as the number of functioning nephrons becomes too low to cope with the phosphate load, despite lowering of the TmP/GFR. Progressive nephron loss and metabolic acidosis further impair $1,25(\text{OH})_2\text{D}$ synthesis. Hypocalcaemia develops and hyperparathyroidism worsens, so by this stage, bone histology in most subjects will show signs of hyperparathyroidism that cannot be reversed by restricting the intake and absorption of phosphate.

Untreated, ~40% of patients reaching established renal failure will have significant parathyroid bone disease, as judged by the presence of marrow fibrosis on bone biopsy. If there is coincident vitamin D deficiency, then florid osteomalacia with hyperparathyroidism can arise.

The interactions between PTH and $1,25(\text{OH})_2\text{D}$ are complex. $1,25(\text{OH})_2\text{D}$ modifies PTH secretion indirectly through its actions on plasma ionized calcium, but it can also act directly on the parathyroids and inhibit the transcription of pre-pro-PTH mRNA. In established renal failure there are fewer receptors for $1,25(\text{OH})_2\text{D}$ in the nuclei of the parathyroid cells, as well as reduced circulating concentrations of $1,25(\text{OH})_2\text{D}$. The loss of this direct effect upon the parathyroids is an important factor in the development of parathyroid hyperplasia. An additional effect of importance is the change in the 'set point' for PTH secretion. This is defined by the concentration of plasma ionized calcium that will suppress PTH secretion by 50% as calculated from calcium infusion experiments. Calcitriol deficiency shifts the set point to the right, that is, a higher ionized calcium has to be achieved before PTH secretion is suppressed.

Aluminium retention

Aluminium retention was recognized as a significant cause of skeletal disease in the late 1970s, when an unusual type of osteomalacia became endemic in haemodialysis units in certain geographical locations, such as Newcastle-upon-Tyne. Aluminium sulphate is added to water as a flocculent during the purification process in order to precipitate organic matter. The quantity of aluminium required for this process varies with environmental factors, so the concentration in tap water may range from 5–30 $\mu\text{mol/L}$. When the dialysate aluminium content is >2 $\mu\text{mol/L}$, there is a substantial transfer from dialysate to plasma. Aluminium is normally excreted by the kidneys, so, in a patient with no renal function, it accumulates in various tissues, including the brain, bone and the parathyroids. The aluminium content of the dialysate can be reduced to <2 $\mu\text{mol/L}$ by reverse osmosis treatment and it is now standard to use dialysates prepared this way. The oral ingestion of aluminium hydroxide as a phosphate-binding agent can also contribute to aluminium loading.

In bone, aluminium is taken up at the bone-osteoid interface and has a marked toxic effect upon osteoblasts, inhibiting the rate of new matrix synthesis and, at higher

doses, inhibiting mineralization. Aluminium can also accumulate in the parathyroids, and patients with severe aluminium-related disease often have low plasma PTH concentrations.

Clinical features

Hyperparathyroidism in the early stages of CKD-MBD is asymptomatic, but with advanced disease, bone pain and proximal myopathy are the predominant symptoms. Because of its insidious onset, it tends to be under-reported and patients may only notice that they have had it when adequate treatment is given and the pain relieved. Fractures are rare in parathyroid disease, but in adolescents and children, horrifying deformities can develop very rapidly as a result of the erosion and weakening of the metaphyses of the long bones. Hyperparathyroidism can impair the effectiveness of erythropoietin therapy.

The full-blown syndrome of aluminium-induced osteomalacia comprises generalized skeletal pain, frequent fractures and disabling proximal myopathy. With very high total body aluminium loads, hypochromic anaemia and dementia ('dialysis dementia') may occur. Because of increased awareness of aluminium, this presentation is uncommon nowadays. Milder degrees of aluminium-related bone disease and a mixed form of aluminium and parathyroid disease are recognized. These forms of disease give rise to fewer symptoms. Aluminium accumulates more readily in the bones of children and adolescents with CKD than in older patients, and also more readily in the bones of patients with CKD secondary to type 1 diabetes.

In recent years, changes in the therapy for CKD-MBD, and changes in dialysis techniques resulting in more active suppression of hyperparathyroidism, have meant that the more florid forms of CKD-MBD have become less common. A consequence of therapeutic measures to suppress hyperparathyroidism has been the emergence of a low turnover state called 'adynamic bone disease'. The clinical sequelae are not clear, but there are suggestions that the risk of fracture is higher and that, in children, growth may be impaired. The strategy of maintaining higher plasma calcium concentrations to suppress hyperparathyroidism also increases the ECF [calcium] \times [phosphate] product, and thus the risk of metastatic or vascular calcification. With an increasing number of older patients being taken onto dialysis programmes, osteoporosis with a high incidence of fractures has also emerged as a significant clinical issue.

Investigations

Biochemical measurements in serum are useful in assessing renal bone disease, but even with the help of radiographic assessments they cannot always provide an unequivocal account of events in bone. CKD-MBD is still, therefore, the commonest indication for bone biopsy, which remains the reference standard for other tests.

Calcium. Plasma calcium concentration (either total or ionized) is usually normal in early renal impairment but

BOX 31.6 Aetiology of hypercalcaemia and hyperphosphataemia in dialysis-treated patients

Hypercalcaemia (>2.7 mmol/L)

- Use of calcium-containing phosphate binders
- Therapeutic vitamin D toxicity
- Severe hyperparathyroidism
- Aluminium bone disease and mixed bone disease
- High dialysate calcium concentrations

Hyperphosphataemia (>2.0 mmol/L)

- Inadequate dialysis regimen
- Inadequate phosphate-binder therapy
- Poor adherence to dietary restrictions
- Therapeutic vitamin D toxicity
- Severe hyperparathyroidism

falls as renal function deteriorates (glomerular filtration rate <30 mL/min). Dialysis partially corrects the hypocalcaemia. In long-term dialysis patients, hypercalcaemia may develop and can arise from one or more of a number of causes (Box 31.6). With the move away from using aluminium-based to calcium-based phosphate binders, the prevalence of hypercalcaemia in dialysis-treated patients has increased.

Phosphate. Plasma phosphate concentration is normal or reduced in early renal disease, but hyperphosphataemia occurs when renal impairment is advanced. In dialysis-treated patients, hyperphosphataemia is inevitable and it is almost impossible to maintain 'normal' plasma phosphate concentrations (0.8–1.4 mmol/L). Phosphate should be kept in the range 1.4–2.0 mmol/L, although even this can be difficult. Phosphate concentrations <1.4 mmol/L can contribute to impaired bone mineralization, whereas at >2.0 mmol/L there is a risk of metastatic calcification (see Chapter 6). Failure to maintain plasma phosphate below this concentration can be due to a variety of reasons (see Box 31.6).

Parathyroid hormone. After secretion of the intact molecule, PTH undergoes cleavage into amino-terminal and carboxy-terminal fragments. The latter are metabolized and excreted predominantly by the kidney, so that immunoassays that are not specific for the intact molecule or the amino terminal are liable to detect high concentrations of circulating PTH, whether or not glandular secretion is increased (see Chapter 6). It may be difficult to interpret a single measurement of plasma PTH in an individual. Nonetheless, in population studies the patients with the most severe parathyroid bone disease have the highest plasma PTH concentrations, irrespective of the assay used, and plasma PTH correlates well with histological indices of bone resorption (and formation). The newer assays for 'intact' PTH perform well in established renal failure but these assays also measure a large PTH fragment (probably PTH 7-84). New two-site chemiluminescence and immunoradiometric assays specific for PTH 1-84 are increasingly used in patients with renal impairment.

FGF23. By the time patients reach dialysis, plasma FGF23 concentrations are extremely elevated (commonly ~1000 times normal). This may reflect delayed clearance as well as increased production of FGF23, but almost all the circulating FGF23 seems to be biologically active. The prognostic value of FGF23 measurements in patients with CKD is beginning to be explored, but at present there is no indication for its routine measurement.

Alkaline phosphatase. The bone isoenzyme is neither dialysable nor filterable by the kidney, so its plasma concentration (and thus activity) is not modified by variations in renal function. Plasma alkaline phosphatase activity is increased when bone turnover is high and, in general, reflects the degree of hyperparathyroidism. If the plasma PTH concentration is clearly elevated, a high ALP is strongly predictive for high-turnover renal bone disease. Aluminium deposition affects osteoblast function so that plasma ALP activity is often normal, or only slightly above the reference range, in aluminium-related disease. The activity of ALP gives no indication of the severity of this type of bone disease.

Other markers of bone turnover. The markers of bone resorption based on urine measurements are obviously of limited use in CKD. Markers such as osteocalcin, hydroxyproline, pyridinoline, deoxypyridinoline and cross-linked C-telopeptide of type I collagen (ICTP) are wholly or partially cleared by the kidneys, so in CKD their plasma concentrations are greatly increased because of accumulation of inactive fragments. In population studies, correlations with histological indices of bone resorption or formation can be demonstrated, but in practice they are of little diagnostic value in individual patients. Concentrations of the formation markers PINP and P1CP are not so affected by CKD, but correlate less well than ALP with histological indices. Tartrate-resistant acid phosphatase 5b is unaffected by CKD and correlates with histological measures of bone resorption. It is elevated in high-turnover bone disease.

Aluminium. Provided that care is taken in the collection and laboratory handling of specimens to avoid contamination, aluminium concentrations in plasma reflect the recent level of exposure to this element. Because of the rapid uptake of aluminium into tissues, plasma concentrations do not reliably indicate the total body burden. The desferrioxamine test was developed in an attempt to measure the total body aluminium load. Desferrioxamine mobilizes aluminium from bone and other tissues, so after its intravenous administration to dialysis patients, plasma aluminium concentrations increase (see Appendix 31.2 for protocol). The increment in plasma aluminium correlates reasonably ($r=0.6-0.7$) with the degree of aluminium loading as assessed by bone biopsy, but the test has insufficient sensitivity and specificity to be used on its own. However, an increment of >7.5 $\mu\text{mol/L}$ in the presence of low plasma PTH is very suggestive of aluminium loading and an indication for bone biopsy to establish the extent of aluminium accumulation. A recommended protocol for surveillance

for aluminium overload in dialysis patients is to measure serum aluminium every 4–6 months, seeking to maintain plasma concentrations at $<1.9 \mu\text{mol/L}$.

Radiology, scintigraphy and densitometry. Radiology is useful in the diagnosis of parathyroid bone disease. Subperiosteal erosions of the phalanges and/or the distal ends of the clavicle are virtually pathognomonic. Other radiographic manifestations of hyperparathyroidism include osteosclerosis (commonly in the spine) and, more rarely, periosteal new bone formation. Aluminium-related disease has no specific X-ray findings, but pathological fractures and radiolucency of bones are typical if there is osteomalacia. Spinal bone density measurements are often increased in hyperparathyroid bone disease and reduced in aluminium-related disease, but these measurements are not of diagnostic value. The scintigraphic changes of hyperparathyroidism add little to the information from other diagnostic modalities.

Bone histology. Bone biopsy after double tetracycline-labelling, with the preparation of undecalcified sections for quantitative histomorphometry, remains an important investigation. Severe hyperparathyroidism can be demonstrated satisfactorily by radiographic and biochemical means, but mineralization defects, low bone turnover and aluminium loading are difficult to quantitate accurately without biopsy. Florid forms of aluminium-related disease are now uncommon, but a mixed form of bone disease with hyperparathyroidism and aluminium retention together, which can only be diagnosed by biopsy, is recognized. Bone histomorphometry permits classification of an individual into one of several diagnostic categories described according to the parameters of turnover (low, normal or high), mineralization (normal or abnormal) and volume (low, normal or high). CKD-MBD remains one of the commonest indications for bone biopsy (Appendix 31.1).

Treatment

Hyperparathyroidism

Vitamin D derivatives. Abnormalities of vitamin D metabolism are critical in the genesis of hyperparathyroidism, and vitamin D, its metabolites or derivatives are successful in reversing hyperparathyroidism at least in the short term. All forms of vitamin D are effective, but forms that require 1α -hydroxylation (such as calciferol or $25(\text{OH})\text{D}$) have to be given in pharmacological doses in order to achieve the desired effects. The 1α -hydroxylated forms, $1,25(\text{OH})_2\text{D}$, alfacalcidol ($1\alpha\text{OHD}_2$) or doxercalciferol ($1\alpha\text{OHD}_2$) are effective at much lower doses and are preferred. Their main advantages are their very quick onset and cessation of action: if vitamin D toxicity occurs with these compounds, it reverses rapidly. Although this is a distinct advantage, toxicity can also arise abruptly. Intestinal calcium and phosphate absorption are increased, the $[\text{calcium}] \times [\text{phosphate}]$ product in plasma increases and, with it, the risk of metastatic and vascular calcification. This can accelerate the decline in

renal function in patients with moderate CKD who have not yet reached dialysis and increases the risk of cardiovascular disease.

How well hyperparathyroidism is suppressed depends mainly on the increment in plasma calcium that can be safely attained. Patients who have high or high-normal plasma calcium concentrations at the outset of therapy are therefore likely to tolerate only low dosages. $1,25(\text{OH})_2\text{D}$ may be given intravenously (usually thrice weekly). The object is to obtain transiently high plasma $1,25(\text{OH})_2\text{D}$ concentrations and to maximize the directly suppressive effects of $1,25(\text{OH})_2\text{D}$ upon PTH synthesis. However, the degree of suppression of hyperparathyroidism obtained by this method remains closely correlated with the increment in plasma calcium. In patients with established hyperparathyroidism, successful treatment with vitamin D derivatives is accompanied by reductions in plasma PTH concentration and ALP activity. Alkaline phosphatase may rise in the first few weeks of therapy, but then declines exponentially with a half-time of around 50 days. As healing takes place, vitamin D toxicity is likely to occur, so dosage reductions are required. All patients taking vitamin D derivatives require monitoring of plasma calcium, phosphate and ALP.

There is increasing use of an emerging group of bioactive vitamin D analogues with supposedly less marked calcaemic activity than calcitriol (e.g. paricalcitol and maxacalcitol). It is hoped that these agents may allow more effective control of parathyroid gland hyperplasia with less hypercalcaemia, although as yet there is little evidence that they are superior to calcitriol, alfacalcidol or doxercalciferol.

Phosphate metabolism. Control of plasma phosphate is important in slowing the development of hyperparathyroidism and is approached by dietary restriction, the use of phosphate-binding agents and adequate dialysis. The concentration of calcium in the dialysate can also be manipulated to influence hyperparathyroidism.

Parathyroidectomy. Parathyroidectomy is very effective in patients with severe hyperparathyroidism and produces a quicker and more complete resolution of parathyroid bone disease than can be achieved with alfacalcidol or calcitriol. With the abrupt cessation of PTH-driven bone resorption after surgery, there follows a period when bone formation remains very rapid, before it, too, slows to match the new, lower resorption rate. During this period, calcium and phosphate are taken up very rapidly from the blood into bone and profound hypocalcaemia (and hypophosphataemia) can occur – the ‘hungry bone syndrome’. This can be so severe as to cause hypocalcaemic fits. Patients undergoing parathyroidectomy are therefore treated pre- and postoperatively with 1α -hydroxylated forms of vitamin D and may require intravenous or oral calcium. After parathyroidectomy, plasma ALP activity may show an early rise but then suppresses rapidly, with a half-time of around 23 days, as bone disease heals. The more rapid suppression of ALP after parathyroidectomy than after alfacalcidol or calcitriol reflects the more rapid reduction in PTH.

Hyperparathyroidism can recur after parathyroidectomy and there is continuing debate over the choice of procedure – subtotal or total parathyroidectomy – and, in the subtotal procedure, about the size of the remnant and whether to leave it in situ or to transplant it to the sternomastoid or forearm muscles. One concern about the rapid suppression of parathyroid hormone, whether by parathyroidectomy or by the use of 1α -hydroxylated vitamin D compounds, is the ease with which aluminium is retained in bone when turnover is very slow. The aluminium–osteomalacia syndrome can certainly be precipitated by parathyroidectomy, hence the importance of making sure the bone is free of aluminium before undertaking this procedure.

Calcimimetic agents. An important development in the treatment of hyperparathyroidism secondary to CKD has been the use of cinacalcet, an allosteric agonist of the ionized calcium-sensing receptor that is expressed on parathyroid cells. With cinacalcet treatment, reductions of ~40% in PTH concentrations are reached after about three months of treatment, and are maintained for at least two years. Reductions in plasma calcium and phosphate of 6–8% are seen and the [calcium] \times [phosphate] product decreases by ~15%.

Aluminium toxicity

Aluminium toxicity is treated by withdrawing all parenteral sources of aluminium and stopping aluminium-containing phosphate binders. The chelating agent desferrioxamine effectively removes aluminium from bone and other tissues. In order to minimize the risk of desferrioxamine-related cerebral, auditory and visual side-effects, and siderophore-mediated opportunistic infections, the chelator should be used at low doses (5 mg/kg) and administered only once weekly – typically for 4–6 months. In patients not requiring dialysis, most of the aluminium–desferrioxamine complex is excreted in the urine and can be monitored by urinary aluminium measurements. Prolonged desferrioxamine treatment can result in iron depletion.

Bone disease after renal transplantation

In the first two weeks following successful renal transplantation, plasma FGF23 concentrations fall very rapidly and $1,25(\text{OH})_2\text{D}$ concentrations rise. Hyperparathyroidism resolves slowly after transplantation and may be expressed as post-transplant hypercalcaemia. Hyperparathyroidism, persistent FGF23 elevation and steroid treatment all contribute to hypophosphataemia in transplant recipients (see Chapter 6). Bone resorption, stimulated by hyperparathyroidism, may be high after transplantation, but corticosteroids, given for immunosuppression, suppress bone formation. Transplant recipients may thus lose bone and become osteoporotic. The suppressive effect of steroids on osteoblasts means that the markers of osteoblast function, ALP activity and osteocalcin are not reliable guides to bone resorption in these circumstances.

BONE DISEASE IN PRIMARY HYPERPARATHYROIDISM

Bone disease was one of the complications of primary hyperparathyroidism that led to the recognition of this disorder in the 1920s, but in Western countries florid parathyroid bone disease is now uncommon. The availability of plasma calcium measurements on modern biochemical analysers has identified a large population with asymptomatic hyperparathyroidism, and early surgery in symptomatic patients probably accounts for the infrequency of severe skeletal manifestations.

Clinical, biochemical and histological features

Since bone is one of the target tissues for PTH, it is not surprising that the bones of almost all patients with hyperparathyroidism show some histological evidence of increased parathyroid activity. In asymptomatic or mildly affected individuals, this amounts only to increases in the extent of active formation and resorption surfaces. This is reflected in modest increases in biochemical markers of bone turnover. Mild primary hyperparathyroidism causes a preferential loss of cortical bone, and may be associated with an increased fracture risk. Both antiresorptive therapy and surgical correction of the disease prevent progressive bone loss.

Much less commonly, patients with primary hyperparathyroidism manifest overt parathyroid bone disease (osteitis fibrosa) characterized by marked increases in osteoclastic bone resorption and increases in bone formation rate, with a resulting increase in osteoid surface extent and marrow fibrosis. Patients experience generalized bone pain, often worse when standing, bone tenderness and proximal myopathy. Radiographic signs can be helpful, subperiosteal erosions being pathognomonic of parathyroid bone disease. Spinal osteosclerosis can occur in primary hyperparathyroidism, but is seen less frequently than in uraemic hyperparathyroidism. Cystic lesions filled with osteoclasts, and marrow fibrosis ('brown tumours'), are a feature of bone disease in severe primary hyperparathyroidism and can be sites of pathological fracture.

The presence of bone abnormalities tends to be associated with more aggressive disease with a shorter duration of symptoms, higher plasma concentrations of calcium and PTH and larger, more rapidly growing brown tumours. Occasionally primary hyperparathyroidism can enter an accelerated phase with a marked increase in the rate of bone resorption and disequilibrium hypercalcaemia (acute parathyroid crisis).

Vitamin D insufficiency occurs commonly in patients with primary hyperparathyroidism, and is probably associated with larger parathyroid adenomas and higher rates of bone turnover. Accelerated hepatic metabolism of $25(\text{OH})\text{D}$ may contribute to the lower vitamin D concentrations in this situation. Maintenance of vitamin D sufficiency may restrain the progression of the parathyroid disease. In patients with mild hyperparathyroidism and low plasma $25(\text{OH})\text{D}$, supplementation with vitamin D

does not usually exacerbate the hypercalcaemia and may, indeed, lower PTH and bone turnover.

Treatment

Most patients with mild primary hyperparathyroidism do not require specific therapy. All patients with severe primary hyperparathyroidism or osteitis fibrosa require parathyroidectomy to correct the disease, but these patients are at risk of developing the 'hungry bone syndrome' with postoperative hypocalcaemia and hypophosphataemia. This can be ameliorated by treatment for 1–2 weeks prior to surgery with low doses of 1 α -hydroxylated vitamin D derivatives. Following successful surgery, plasma calcium concentrations fall, but there can be secondary hyperparathyroidism as the remaining normal parathyroid glands respond to the fall in plasma calcium. The changes in plasma calcium following surgery are therefore a better immediate guide to the success of the procedure than changes in plasma PTH concentration. Nonetheless, intraoperative measurement of PTH can provide useful information as to whether the abnormal parathyroid tissue has been resected – a fall of >50% in the PTH concentration within 20 min of parathyroid gland removal is highly predictive of long-term surgical cure. Plasma ALP may flare in the 1–2 weeks following surgery, and then falls in a similar manner to that seen after parathyroidectomy in uraemic hyperparathyroidism. The biochemical markers of bone resorption fall early after successful surgery.

PAGET DISEASE OF BONE

Epidemiology

Although Paget disease of bone is, next to osteoporosis, the metabolic bone disorder most frequently encountered in general medical practice, it is in many ways quite mysterious. The disease is characterized by extreme metabolic disturbance of one (monostotic) or more (polyostotic) bones in the skeleton, while the uninvolved bones are quite normal. This patchy, asymmetrical distribution of disease within the skeleton is unexplained, as is its striking geographical distribution. It is most common in people of European origin living in temperate zones and is common in the UK, the USA, Australia, New Zealand and western Europe. It has been described in African-Caribbean peoples living in the UK and in African-Americans, but it is rare in groups native to areas where it is common among whites, such as native North Americans, Australian aborigines and the Maori of New Zealand. Even in countries where the disease is widespread, there are well-documented variations in prevalence. For example, in the UK, the prevalence in parts of Lancashire is two to three times greater than elsewhere. Families with many members affected, and a dominant pattern of inheritance, are not uncommon.

Aetiology

The cause of Paget disease is unknown, but both genetic and environmental factors appear to be important. In about

a third of patients with a positive family history (and a much lower proportion of patients without a positive family history), Paget disease is associated with mutations in the gene *SQSTM1*. This gene encodes the p62 protein, which forms part of the scaffold linking the RANK receptor on osteoclasts to the nuclear transcription factor NF- κ B. The majority of Paget disease-associated mutations identified to date (currently numbering ~30) have been in the ubiquitin-binding domain of the molecule. Patients with *SQSTM1* mutations tend to have disease that is more extensive, more symptomatic and is of earlier onset than in patients without such mutations. This is particularly the case with mutations that result in truncation of the p62 protein. The commonest *SQSTM1* mutation is P392L which is found all around the world. Haplotype evidence suggests this is a founder mutation distributed by European colonists. The rare dominantly inherited syndrome of inclusion-body myopathy, Paget disease and frontotemporal dementia (OMIM 605382) is associated with mutations in the *VCP* gene, encoding the valosin-containing protein. This protein is also a component of the scaffold linking RANK to NF- κ B.

Genome-wide association studies of patients negative for *SQSTM1* mutations have indicated that polymorphisms at a number of genetic loci are statistically associated with Paget disease. Several of the candidate genes encode proteins known to be important in osteoclast physiology, such as M-CSF, a cytokine essential for osteoclast differentiation, DC-STAMP, a cell surface protein essential for the formation of multinucleated osteoclasts and RANK, a receptor essential for osteoclast differentiation and bone resorption.

Recent epidemiologic data indicate that in the past 40 years, there has been a significant reduction in the prevalence of Paget disease in high-prevalence regions such as the UK and New Zealand, and that extensive polyostotic involvement has become less common. These observations suggest that in addition to genetic influences an important environmental factor is involved in the causation of Paget disease. Research has concentrated on a potential viral cause as the environmental agent. Both measles and canine distemper virus have been suggested as culprits, but the laboratory evidence is inconclusive.

Natural history

Many people have the disease without symptoms, as shown by the frequent incidental finding of either an isolated elevation in plasma ALP activity in 'routine' blood tests or pagetic changes in bone when radiographs are taken for other indications. Radiographic surveys have shown that in endemic areas of the world, the prevalence increases with age, and suggest that disease before the age of 40 is rare. It appears to arise more or less simultaneously at several skeletal sites. In some patients, distinct phases can be discerned. Bone lysis and progression of lytic lesions may predominate, particularly in the skull and long bones: this is presumed to be early disease. In the later stages, in some patients, osteoclast activity dies out, the bones become intensely sclerotic and ultimately osteoblast activity declines too, leaving abnormal but metabolically quiescent bone. What proportion of patients go through this cycle and how long it takes are not known. The majority

of patients have mixed disease with similar increases in both resorption and formation of bone. Observations of plasma ALP activity in untreated subjects show an unpredictable variation from year to year, with an overall tendency to rise with time until it becomes stabilized within a range characteristic for an individual.

Pathology

Paget disease is characterized by massively increased turnover in affected bones. It is thought to be primarily a disorder of osteoclasts. The number and size of the osteoclasts are greatly increased and they contain numerous large nuclei. Resorption surfaces are increased in extent. The coupling of bone resorption to formation is, however, more or less maintained, and many of the histological features of Paget disease relate to the massive increase in bone formation. Osteoblast numbers, the extent of bone-forming surfaces and the rate of new bone formation are all increased. At very high rates of bone formation, collagen fibres are not laid down in the usual orderly lamellar pattern, but more randomly, resulting in new bone of abnormal structure (visible as a mosaic pattern under polarized light), known as 'woven bone'. This bone has a larger volume than lamellar bone, hence pagetic bone shows an increase in trabecular bone volume. Collagen fibres are also secreted by osteoblasts into the marrow space, giving rise to marrow fibrosis.

Clinical features

Bone pain is a common feature. Pagetic bone pain is characteristically chronic and persistent rather than episodic, worse at rest and relieved by movement. It responds rapidly to antipagetic therapy. Because pagetic bone is soft and readily deforms, secondary osteoarthritis is very common. This is particularly so in the hip joints (from pelvic or femoral involvement). Pathological fractures of long bones (particularly in areas where lysis dominates) are common. Pagetic bone is very vascular and the high blood flow causes an elevation of skin temperature over the affected bone. In severe polyostotic disease in the elderly, the quantity of blood shunted through bone may precipitate heart failure. Spinal disease can cause a reversible myelopathy through a vascular steal phenomenon. Involvement of the otic capsule causes deafness. The most feared complication is the development of osteosarcoma, but fortunately this is rare. Osteosarcoma appears to arise from somatic mutations.

Investigations

Radiology

The radiographic appearances of bone reflect the histological events. There are areas of focal bone resorption (lytic lesions). In long bones and in the skull, flare-shaped lytic fronts can progress through bone: typical rates of progression are 0.8–1.2 cm/year. The increased bone formation is reflected by loss of the normal trabecular pattern, sclerosis and an overall enlargement of bone dimensions. Deformity, arthritis and fracture are common secondary events visible

on radiographs. Incomplete cortical 'fissure' fractures of the convex surface of deformed long bones often precede completed fractures. Isotope scanning is the best method of determining the extent of the disease.

Biochemical tests

Bone turnover markers. Biochemical tests are very useful in monitoring the progress of Paget disease and its response to therapy. The markers in use are not specific to the disease, but are simply markers of increased rates of bone formation and resorption. The concentrations of plasma ALP and other markers reflect both the extent and the activity of the disease. Hence the greatest values are seen in patients with widespread active polyostotic disease. In such patients, the turnover markers are elevated to a greater extent than in any other metabolic bone disorder. In patients with monostotic disease, bone turnover markers need to be interpreted with caution; although increased for that individual, they may not be outside the reference range.

All bone turnover markers are increased proportionately apart from osteocalcin, which is not increased to the extent expected from ALP measurements. Because of its rapid turnover, much of pagetic bone is 'young' and thus has a greater proportion of the linear, non-isomerized α -C-telopeptide collagen cross-links, and a smaller proportion of the β -isomerized C-telopeptides typical of older bone. This is reflected in the results of assays for α -CTX and β -CTX – patients with Paget disease having a greater than usual proportion of the former.

Alternative methods of monitoring disease activity can be employed. Over the lower limb and the forearm, skin temperature is a useful guide to disease activity, provided the disease is unilateral so that there is a normal limb for comparison. The skin temperature at the same point on opposite limbs in normal subjects does not differ by more than 0.5°C. The skin overlying a tibia affected by lytic Paget disease can be up to 4°C warmer than the unaffected limb opposite. Alternatively, the uptake of isotope on a bone scintiscan gives a measure of disease activity.

Since bone resorption and formation are closely linked, there is, not surprisingly, a strong correlation between plasma ALP activity and bone resorption markers in the untreated state. It is arguable as to whether, in the presence of symptoms, signs and radiographic evidence of Paget disease, it is necessary for diagnostic purposes to measure resorption markers if the plasma ALP is elevated and other liver-derived enzymes are normal. The smaller coefficient of variation in ALP measurements at times when disease activity is stable (~10%), compared with that of urine resorption markers, is an additional argument in favour of using the former as the main index of disease activity. A potential advantage of measuring resorption markers is the earlier detection of relapse after treatment, but the greater imprecision of the resorption markers means that in practice there is little advantage over measuring formation markers.

In response to treatment with potent bisphosphonates, the bone markers that appear to perform best (i.e. show the greatest changes) are total ALP, bone ALP and PINP. The latter probably performs best in terms

of pre-treatment values being clearly elevated in disease of limited extent, and showing the greatest changes with treatment and on relapse of disease. Using the same criteria, the best performing resorption marker is the urine N-telopeptide/creatinine ratio.

Plasma and urinary calcium. Plasma and urinary concentrations of calcium (and phosphate) are usually normal in Paget disease, but patients with very extensive and active disease are prone to develop hypercalciuria and even hypercalcaemia if immobilized. This results from increased bone resorption and a relative fall in bone formation, detectable by an increase in bone resorption markers and a fall in ALP activity.

Responses to treatment

There is no cure for Paget disease, but the metabolic activity of the affected bone can be reduced by the use of inhibitors of bone resorption – nowadays bisphosphonates are used almost exclusively. The main indication for treatment is bone pain. In practice, it can be difficult to distinguish pagetic from arthritic pain, so a trial of therapy may be given. Prompt (and prolonged) relief after treatment suggests that Paget disease was the main cause of pain. Myelopathy secondary to vascular steal is an indication for immediate treatment. Bisphosphonates

are commonly give in preparation for joint replacement surgery, where one of the bones concerned is involved with Paget disease, in the hope that this will reduce the vascularity of the bone and make the surgery easier. There is, as yet, no clear evidence that aggressive treatment to reduce bone turnover to normal prevents fracture or the development of late complications such as deafness or arthritis, though it is probable (comparing radiographic series from the pre-bisphosphonate era and now) that deformity can be prevented.

In response to bisphosphonate treatment, there is an early fall in urine resorption markers reflecting the reduction in osteoclastic activity. This rapid fall is complete by four days after intravenous treatment. Bone formation falls secondarily to this, so that the decline in plasma ALP lags behind that of bone resorption markers by several months. Ultimately, a new steady state is reached where resorption and formation are matched (Fig. 31.9). When the treatment is with oral bisphosphonates, a similar pattern is seen. Both bone resorption and bone formation markers fall in a monoexponential manner with a half-time of around two weeks, but the response of the formation markers lags behind by two to three weeks.

During the period early after initiation of therapy where resorption has fallen but formation remains high, there is a considerable uptake of calcium and phosphate into bone, thus some degree of secondary hyperparathyroidism is

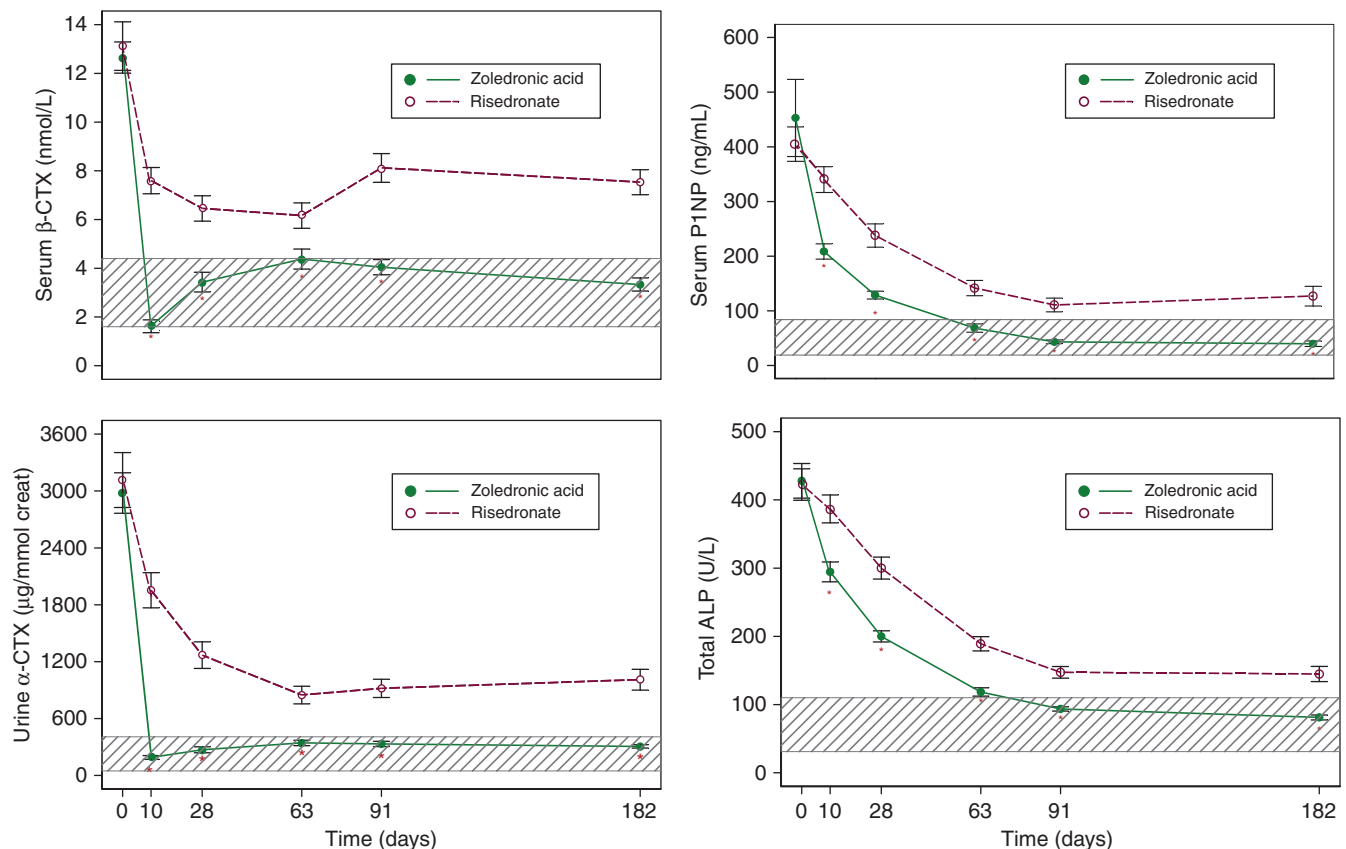


FIGURE 31.9 ■ The treatment of Paget disease. Effects of a single infusion of zoledronate 5 mg or oral risedronate 30 mg/day for two months on bone turnover in patients with Paget disease. The very rapid drop in bone resorption markers (serum β -CTX and urine α -CTX) is followed by a slower fall in the formation markers (serum P1NP and ALP). Zoledronate produces a more profound and sustained suppression of bone turnover. Hatched areas indicate normal reference intervals. (Based on data from Reid et al. 2005 in *N Engl J Med* 353: 898–908.)

usual. Following intravenous bisphosphonates, this is evident some 2–4 weeks after the infusions but it is not of any great consequence in most patients. Plasma osteocalcin concentrations may actually increase by 30–40% within the first two weeks of treatment, probably because the secondary hyperparathyroidism stimulates an increase in $1,25(\text{OH})_2\text{D}$ synthesis and this, in turn, stimulates osteocalcin production. Osteocalcin concentrations do subsequently decline with the fall in bone formation. Because of these disparate responses and the poor correlation with ALP in the untreated state, plasma osteocalcin measurements are not used in the management of Paget disease.

Treatment with calcitonin produces reductions in both plasma ALP activity and urinary hydroxyproline excretion to an average of 50% of the pretreatment values, within six months of starting treatment. Continued use of calcitonin produces no further improvement – the ‘plateau’ phenomenon – and discontinuation of treatment is followed by prompt relapse.

The bisphosphonates have made the use of calcitonin and other treatments obsolete, since they are far more effective and have fewer side-effects. A large number of bisphosphonates have been used. The more recently developed, such as zoledronate, are substantially more potent and have a long duration of action. Patients whose turnover markers suppress to the lowest values are likely to remain in biochemical ‘remission’ the longest. For example, if plasma P1NP is suppressed to $<40\ \mu\text{g/L}$ six months after treatment with intravenous zoledronate, then $>90\%$ of patients will remain in biochemical remission 6 years later. If P1NP suppresses only to the 60–80 $\mu\text{g/L}$ range, then only $\sim 60\%$ will remain in remission. If treatment is discontinued, all patients seem ultimately to relapse – although in patients with limited disease, treatment with potent bisphosphonates can produce remissions of ten years or more.

Relapse is first detectable biochemically by an increase in markers of bone resorption. Retreatment when a relapse has occurred produces a further suppression of bone turnover markers, but it is proportionately smaller, so the previous nadir is reached but concentrations are not usually reduced much further. The bisphosphonates, though effective orally, are poorly absorbed from the gut, so short intravenous regimens are frequently employed. The more potent amino-bisphosphonates (such as pamidronate, ibandronate and zoledronate), when given intravenously for the first time, often provoke an acute febrile reaction.

Etidronate, the first bisphosphonate to be developed, has a toxic effect upon osteoblasts and impairs mineralization when given in high doses ($>10\ \text{mg/kg}$ per day) for >3 months. Etidronate also has a much greater effect on the TmP/GFR than the other bisphosphonates, so that hyperphosphataemia is not uncommon during treatment. It is little used nowadays as the newer bisphosphonates are significantly more effective and longer lasting.

BONE TURNOVER AND BONE DISEASE IN CHILDREN

In infants, children and adolescents, bone turnover, as assessed by biochemical markers, is at much higher levels than in healthy adults. In general, the concentrations

TABLE 31.5 Bone turnover abnormalities in paediatric bone diseases

Disorder	Abnormalities
Prematurity	Bone ALP and osteocalcin relatively reduced, but increase after several weeks
Malabsorption/ malnutrition	All turnover markers relatively reduced
Growth hormone deficiency	Relative reduction in formation markers – which increase with growth hormone treatment
Glucocorticoid use	Relative reduction in both formation and resorption markers
Vitamin D deficiency	Increased ALP and P1NP, but osteocalcin relatively low: resorption markers increased

reflect growth velocity. [Figure 31.3](#) illustrates age-related changes in ALP and osteocalcin: the mean values at peak growth velocity are 3–5-fold higher than mean adult values. For P1NP, mean values at peak growth velocity are 6–10-fold higher than mean adult values. Turnover markers cannot accurately predict the accrual of bone during growth. The interpretation of bone marker data in individual children is difficult as the growth rate, nutritional status, sex, age, pubertal stage and renal function of the child all need to be taken into account.

There is relatively little information about very early life, but in healthy infants the urine marker of bone resorption NTX is low at birth, increases dramatically in the first ten days of life, remains very high for the next three months of life and then decreases to values similar to those at birth and one year of age.

Because new bone formed during growth is necessarily immature, the α -CTX isomer of the C-telopeptide predominates when this index is measured. Abnormalities in bone markers can be seen in various acquired and genetic disorders that affect bone directly or indirectly, and these are occasionally of diagnostic value ([Tables 31.5](#) and [31.6](#)).

GENETIC BONE DISEASES

There are numerous genetic bone diseases. The discussion below refers only to those conditions not covered above, for which there is a clear role for the clinical biochemistry laboratory in diagnosis (see [Table 31.6](#)).

Osteogenesis imperfecta

The term osteogenesis imperfecta (OI) encompasses a group of genetic disorders usually defined by recurrent fractures, low bone mass and skeletal fragility. Most cases are dominantly inherited and result from mutations in the genes encoding type I collagen (*COL1A1* and *COL1A2*). The condition is not synonymous with type I collagen abnormalities, as a number of genetic disorders affecting type I collagen or its post-translational modification that have skeletal phenotypes are not usually included in the

TABLE 31.6 Biochemical abnormalities in genetic bone diseases

Disease	Gene	Effects	Biochemical findings
Osteogenesis imperfecta*			
Haploinsufficiency	<i>COL1A1</i>	Insufficient type I collagen produced	Low plasma P1NP and P1CP concentrations
'Type VI'	<i>SERPINF1</i>	?	Low plasma concentrations of pigment epithelium-derived factor (PEDF)
Bruck syndrome and Kuskokwim disease	<i>PLOD2</i> <i>FKBP10</i>	Impaired telopeptide cross-linking	Reversal of usual pyridinoline: deoxypyridinoline molar ratio in urine from ~9:1 to ~1:7
Ehlers–Danlos syndrome*			
kyphoscoliotic type (VIA)/Nevo syndrome	<i>PLOD1</i>	Lysyl hydroxylase deficiency	Reversal of usual pyridinoline: deoxypyridinoline molar ratio in urine from ~9:1 to ~1:7
Osteopetrosis			
Autosomal recessive (severe)	<i>TCIRG1</i>	Impaired acidification of extracellular compartment of the osteoclast	Increased TRAP
Autosomal recessive or dominant (severe)	<i>CICN7</i>		Increased TRAP and BB fraction of creatine kinase
Autosomal recessive with renal tubular acidosis (RTA)	<i>CAII</i>		RTA
Autosomal recessive (intermediate) Pycnodysostosis	<i>PLEKHM1</i> <i>CTSK</i>	Impaired vesicular trafficking Cathepsin K deficiency	Increased TRAP Cathepsin K concentrations low, ICTP/CTX ratio increased
Osteosclerosis			
Progressive diaphyseal dysplasia	<i>TGFB1</i>	Mutations affect latency-associated peptide of TGFβ1 ↑TGFβ1 activity	ALP and other bone turnover markers modestly increased
Sclerosteosis and Van Buchem disease	<i>SOST</i>	Inactivating mutations in sclerostin	ALP and other bone turnover markers modestly increased
Juvenile Paget disease	<i>TNFRSF11B</i>	Osteoprotegerin deficiency	ALP and other bone turnover markers very high
Familial expansile osteolysis	<i>TNFRSF11A</i>	Activating mutations of RANK	ALP and other bone turnover markers increased significantly
Polyostotic fibrous dysplasia	<i>GNAS1</i>	Activating somatic mutations in G _s α	ALP and other bone turnover markers increased significantly

*Note: Only variants with distinctive biochemical findings are listed. Please refer to text for explanation of abbreviations.

OI rubric. These include Caffey disease, and the type VI and VII variants of the Ehlers–Danlos syndrome. In addition, some genetic disorders of skeletal fragility included in classifications of OI are the consequence of mutations in key osteoblast genes (*LRP5*, *WNT1* and *SP7*) that code for proteins concerned with matrix homeostasis, and are not directly related to collagen metabolism and matrix structure.

More than 200 mutations in the collagen genes *COL1A1* and *COL1A2* that cause OI have been identified. All the mutations result in decreased synthesis and secretion of normal type I procollagen. When no abnormal procollagen is synthesized, the phenotype is generally mild: this is the commonest form of the disease, arising where a mutation in one of the *COL1A1* genes produces a null allele (haploinsufficiency). When abnormal molecules are synthesized in addition to normal type I collagen, the phenotype can range from very mild to lethal, depending on the nature and location of the mutation.

The diagnosis of OI is predominantly based on clinical assessment (including family history) and imaging. Biochemical tests are helpful with excluding differential diagnoses, but with a few exceptions are not generally specific. Ultrasound and radiographic examinations are useful for intrauterine diagnosis of severely affected fetuses. The quantity and some structural characteristics of type I collagen produced by cultured skin fibroblasts can be assessed in vitro.

The usual biochemical indices of metabolic bone disease are normal in osteogenesis imperfecta, though resorption markers are often increased when the patient is immobilized. In patients with haploinsufficiency, the bone formation markers plasma procollagen C-terminal peptide and procollagen N-terminal peptide (which are derived from type I collagen) are often relatively low compared to the other formation markers such as bone ALP.

The genetic bases of a number of *recessively-inherited* OI variants have been identified. The majority of the genes involved code for proteins that regulate the post-translational modification of type I collagen, and there are some instances where specific biochemical tests can help. In patients with Bruck syndrome (OI with contractures), where there is defective cross-linking of collagen (owing to mutations in *FKBP10* or *PLOD2*) the usual ratio of pyridinoline to deoxypyridinoline in urine is reversed. Another recessive OI variant results from mutations in *SERPINF1* that encodes pigment epithelium-derived factor (PEDF), a secreted glycoprotein of as yet uncertain function in bone. Pigment epithelium-derived factor can be measured in normal serum and is undetectable in the serum of patients with this OI variant.

Bisphosphonate treatment is now widely used in children with more severe forms of osteogenesis imperfecta. During growth, osteoclasts remove bone from the endosteal surface of the diaphysis and the primary

spongiosa in the metaphysis, as part of the modelling process. Bisphosphonates inhibit this osteoclastic resorption, with the end result that cortices are thicker and trabecular number is increased. Bone pain is reduced, the bones are stronger and fracture less frequently. The benefit of bisphosphonates in children and adults with milder forms of osteogenesis imperfecta is less certain.

High bone mass

High bone mass disorders are classified according to whether there is excessive bone formation (osteosclerosis) or a failure of bone resorption (osteopetrosis).

Osteopetrosis

Osteopetrosis is further categorized according to whether there are normal or increased numbers of osteoclasts: osteoclast-rich (the majority of cases) or osteoclast-poor. Osteoclast-rich osteopetrosis arises because of mutations in a number of genes encoding proteins important in generating and excreting hydrogen ions onto the bone surface sealed beneath the osteoclast. The genes usually involved are *TCIRG1*, *CICN7*, *PLEKHM1* or *CAII*. Osteoclasts are typically abundant, but their capacity to resorb bone is impaired. Severe osteopetrosis typically becomes manifest during infancy. The bone marrow space is occluded so the infant develops leukoerythroblastic anaemia with bruising and bleeding and extramedullary haematopoiesis. Growth is poor and blindness and deafness occur because of cranial nerve entrapment. The radiographic features are diagnostic. The plasma calcium concentration may vary and seems to reflect the dietary calcium intake. Restriction of dietary calcium may precipitate tetany, presumably because of defective osteoclastic bone resorption. Urinary calcium excretion is often low. Plasma total and tartrate-resistant acid phosphatase activities are increased, although ALP is normal. This condition can be cured by successful bone marrow transplantation, which permits colonization of the host marrow by normal osteoclast precursors.

Carbonic anhydrase II deficiency, resulting from mutation in the *CAII* gene, is a congenital disorder comprising osteopetrosis, renal tubular acidosis and cerebral calcification. Although haematopoiesis is not affected, cranial nerve entrapment may be a feature. The laboratory findings include a metabolic acidosis and greatly reduced activity of the enzyme carbonic anhydrase II in erythrocytes. This enzyme is widely distributed and is found, among other places, in normal osteoclasts.

The osteoclast-poor forms arise from inactivating mutations in *TNSF11* (encoding RANKL) or *TNFRSF11A* (encoding RANK). Children with the latter also have hypogammaglobulinaemia and commonly develop severe but transient hypercalcaemia after bone marrow transplantation.

Progressive diaphyseal dysplasia

This osteosclerotic condition results from mutations in the gene encoding TGF β . It is characterized by new bone formation, affecting predominantly the diaphyses of the long bones and the skull. The clinical features may

include difficulty in walking, muscle wasting and leg pain. Laboratory findings of note are increased plasma ALP activity and other bone turnover markers.

Familial or idiopathic hyperphosphatasia (juvenile Paget disease)

This is a very rare, recessively inherited disease that results from deletion of, or inactivating mutations in, the gene for osteoprotegerin (*TNFRSF11B*). The onset is in infancy with skeletal abnormalities including dwarfism, bone fragility, a large head and anterior bowing of the limb bones. The bone shows greatly increased cellular activity. The characteristic biochemical changes are very high levels of plasma ALP activity and all other bone turnover markers.

Other disorders

Familial expansile osteolysis and related disorders

These rare, dominantly inherited conditions result from activating mutations in the signal peptide domain of RANK. Painful, progressively worsening expansile lytic lesions develop in bone. The characteristic biochemical changes are a sustained rise in plasma ALP activity and all other bone turnover markers.

Fibrogenesis imperfecta ossium

Fibrogenesis imperfecta ossium is a disorder of bone matrix in which collagen is laid down in a haphazard fashion. The onset is after the age of 50 years and it presents with severe bone pain and tenderness. The plasma ALP activity is consistently elevated.

Polyostotic fibrous dysplasia

Polyostotic fibrous dysplasia is characterized by well-circumscribed areas of marrow fibrosis that replace bone and give the radiographic appearance of cysts, through which pathological fractures may occur. Plasma ALP and urinary hydroxyproline are variably increased, depending on the extent of the disease. The major biochemical interest in this disorder is its association with endocrine diseases such as thyrotoxicosis, acromegaly, Cushing syndrome, precocious puberty and hypophosphataemic osteomalacia (McCune–Albright syndrome). The endocrine disorders are the consequence of constitutive (non-ligand-dependent) activation of the adenylate cyclase second messenger system. They arise from somatic mutations in the gene for the α subunit of the stimulating guanine binding protein ($G_s\alpha$), which forms a critical part of the adenylate cyclase activating system, linking the receptor to the second messenger.

CONCLUSION

Bone is a metabolically active tissue. In mature bone, there is a continuous cycle of bone resorption and replacement. Osteoblasts are responsible for the synthesis

of bone matrix and osteoclasts for bone resorption. The important role of osteocytes is increasingly recognized. Normal bone formation, metabolism and repair depend on the coordinated activity of these cells and the integrity of the homeostatic mechanisms for calcium and phosphate, which primarily involve parathyroid hormone, FGF23 and vitamin D.

Metabolic bone disease can be due to intrinsic abnormalities of the cellular elements of bone or to abnormalities of the homeostatic mechanisms for calcium and phosphate.

The commonest metabolic bone diseases are osteoporosis – a generalized reduction in bone density due to an imbalance between the rates of bone formation and resorption, which can have several causes, of which oestrogen deficiency in postmenopausal women is particularly important; osteomalacia, a defect of mineralization related either to a lack of calcium or phosphate or to defective osteoblast function; CKD-MBD, a complex bone disease occurring in patients with chronic kidney disease, and Paget disease, a condition of unknown aetiology characterized by greatly increased osteoclastic activity and disordered new bone formation.

Biochemical measurements are of considerable importance in the diagnosis of metabolic bone diseases and in the assessment of the response of patients to treatment.

Further reading

Kanis JA. *Pathophysiology and treatment of Paget's disease of bone*. 2nd ed. London: M Dunitz; 1998.

Definitive account of Paget disease.

Rosen CJ, editor. *Primer on the metabolic bone diseases and disorders of mineral metabolism*. 8th ed. Ames, IO: Wiley/American Society of Bone and Mineral Research (ASBMR); 2013.

A succinct and authoritative account of the current state of knowledge in the fields of bone biology, calcium metabolism and metabolic bone disease, written by the leaders in the respective fields.

Seibel MJ, Robins SP, Bilezikian JP, editors. *Dynamics of bone and cartilage metabolism*. 2nd ed. London: Academic Press; 2006.

Good background information on bone biochemistry.

Smith R, Wordworth P. *Clinical and biochemical disorders of the skeleton*. Oxford: Oxford University Press; 2005.

An excellent reference for disorders of the skeleton associated with known biochemical abnormalities and genetic disorders.

APPENDIX 31.1: INDICATIONS FOR DIAGNOSTIC TRANSILIAC BONE BIOPSY

- Excessive skeletal fragility in unusual circumstances (e.g. in a young adults).
- When a mineralizing defect is a diagnostic possibility.
- Characterization of bone lesion in CKD-MBD.
- Diagnosis and assessment of response to treatment in vitamin D-resistant osteomalacia and related disorders.
- When a rare metabolic bone disease is suspected.

To get maximum information from a biopsy, fluoro-chrome labelling of bone is advised. Tetracycline (250 mg q.d.s.) or demeclocycline (150 mg q.d.s.) should be administered by mouth for two days, 18 and 17 days before the biopsy, and again five and four days before the biopsy. Tetracyclines are taken up into the mineralization front and 'double labelling' permits calculation of bone formation and mineralization rates. The biopsy should be collected into phosphate-buffered formalin or 70% alcohol and not decalcified. An experienced bone histomorphometrist is needed to interpret the biopsy.

APPENDIX 31.2: PROTOCOL FOR DESFERRIOXAMINE TEST IN DIALYSIS PATIENTS

Desferrioxamine (DFO) 5 mg/kg body weight is administered i.v. (in 100 mL 5% dextrose during the last 60 min of a dialysis session in haemodialysis patients). The plasma aluminium concentration is determined before this session and before the subsequent session (i.e. approximately 44 h after DFO administration).

An increment in plasma aluminium >7.5 µmol/L suggests aluminium accumulation in bone. When coupled with a low concentration of PTH (<15 pmol/L with an intact PTH assay), it is indicative of a high total aluminium burden, and has a sensitivity of 87% and specificity of 95% for the diagnosis of aluminium-related bone disease.

If the plasma PTH is >65 pmol/L, patients are not prone to the effects of aluminium at the bone mineralization front.

Biochemistry of articular disorders

Jeremy G. Jones

CHAPTER OUTLINE

INTRODUCTION 636

THE ARTICULAR SYSTEM 636

DISORDERS OF THE ARTICULAR SYSTEM 637

Osteoarthritis (OA) 637

Inflammatory arthritis 637

The connective tissue diseases 638

Aches and pains 638

Crystal arthritis 639

ARTICULAR INVOLVEMENT IN ENDOCRINE AND METABOLIC DISEASES 643

Diabetes mellitus 643

Other endocrine disorder 643

Haemochromatosis 643

Alkaptonuria 643

LABORATORY TESTING IN ARTICULAR DISEASE 643

Anaemia in rheumatoid arthritis 643

The acute phase response 644

Examination of synovial fluid 644

Rheumatoid factor 644

Other autoantibody tests 644

CONCLUSION 645

INTRODUCTION

While biochemical abnormalities themselves are responsible for few articular disorders, clinical biochemistry plays an important part in the diagnosis, treatment and monitoring of many of these conditions.

This chapter starts with a description of the articular system and its diseases, which will put the role of biochemistry in these conditions into perspective.

Crystal synovitis, as seen in gout and pseudogout, is the only common articular condition caused by an underlying biochemical abnormality. This is described in some detail along with examination of synovial fluid, the one biochemical investigation that is primarily the province of the rheumatologist.

Some diseases with a primary biochemical abnormality like diabetes, alkaptonuria, endocrine conditions and haemochromatosis have articular manifestations and also warrant discussion.

Immunological abnormalities are at the root of an important group of rheumatic diseases, the connective tissue diseases. Autoantibody formation is a major disease manifestation and, in some hospitals, tests to identify and measure autoantibodies are performed in the biochemistry laboratory. These tests are described briefly.

THE ARTICULAR SYSTEM

The joint is the pivotal structure of the articular system. Joints can be considered as discontinuities in the skeleton that permit controlled mobility. They have different structures depending on their functional requirements.

When no movement is needed, the joint is bound together by tough fibrous tissue (e.g. the skull 'sutures'). In cartilaginous joints, the bone ends are joined by compressible fibrocartilage and reinforced by a surrounding tough fibrous tissue (as in the symphysis pubis and the manubriosternal joint) and permit only a limited amount of movement. This type of joint tends to be located centrally. If a moderate or wide range of movement is required, a space must exist between the bone ends, forming a discontinuous 'synovial' joint. Most of the peripheral joints fall into this category.

The most important joints from a clinician's point of view are the peripheral ones, which enable our arms and legs to move through such a wide range, our hands to be endowed with such dexterity and strength and our feet to carry us so uncomplainingly when we stand, walk, run and jump.

While synovial joints have the same basic structure and physiology, individual joints have evolved differently, depending on their situation and what is expected of them. For example, the knee, ankle and finger joints, which move mainly in one plane, are 'hinge' joints, while the hip and shoulder, which move in all directions, are ball and socket joints. Of course, to carry out even the simplest of tasks, like opening a door or walking up a step, we have to coordinate the movement of several joints, not to mention contracting and relaxing the muscles on either side of the joints.

The archetypal synovial joint is an enclosed space with a negative pressure (Fig. 32.1). The bone ends that move against each other (the articular surfaces) are covered with articular cartilage. This is made up of proteoglycans and collagen, combined in such a way as to allow it to absorb

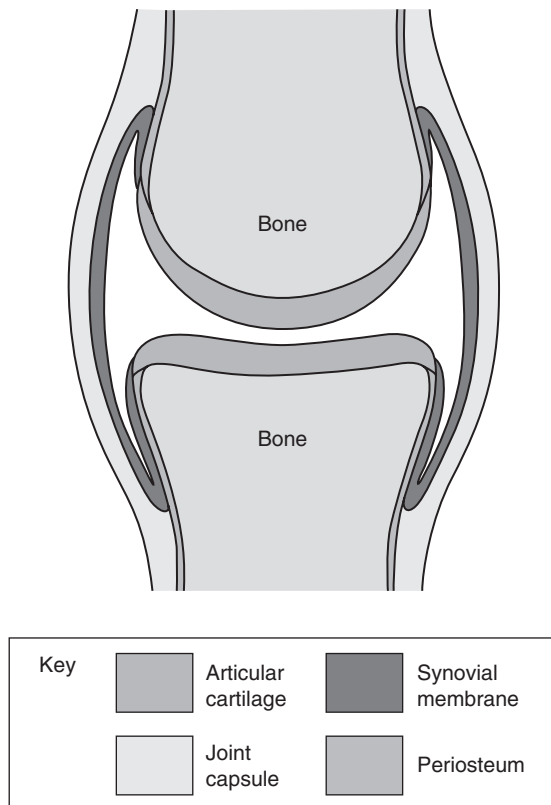


FIGURE 32.1 ■ A typical synovial joint.

huge forces of pressure like a shock absorber, while providing a shiny surface for smooth, low-friction movement. A healthy joint is lubricated by a small amount of synovial fluid (0–4 mL). This is an ultrafiltrate of plasma with additional components secreted by the synovium. The most important of these is hyaluronate, a linear repeating disaccharide with a molecular weight of some 10^7 Da. This provides the fluid with its viscoelastic properties without which the cartilage would fail and smooth movement would be impossible. Cartilage and synovial fluid together maintain coefficients of friction of <0.02 .

The bone ends are enclosed by the tough joint capsule which is lined by the synovial membrane, a structure only a cell or two thick (Fig. 32.1). Synovial cells phagocytose intra-articular debris, secrete many components of synovial fluid and possess immunological functions.

While we talk of ‘articular disorders’, it is not really possible to consider the joint in isolation. Each joint depends on other structures like ligaments, tendons, muscles and bone for its function and stability. Bone and muscle disease are dealt with in separate chapters.

DISORDERS OF THE ARTICULAR SYSTEM

There are over 100 rheumatic disorders with different underlying causes, different treatments and different outcomes. These conditions can be broadly divided into non-inflammatory diseases, of which osteoarthritis is the most important, and inflammatory conditions such as rheumatoid arthritis (RA), ankylosing spondylitis and systemic lupus erythematosus (SLE or lupus).

Osteoarthritis (OA)

The most common form of articular disease is osteoarthritis, in which the articular cartilage becomes fissured and gradually wears away. The joint tries to heal itself by forming bony out-growths on the sides of the joints (osteophytes). Trauma to a joint predisposes to OA, as do some rare biochemical disorders like alkaptonuria. However, in most cases, apart from an inherited tendency and increasing frequency with age, the cause cannot be identified. For many years, OA was thought to be due to degeneration or ‘wear and tear’. Recently, we have appreciated that a degree of inflammation is often present if sought and that calcium pyrophosphate and basic calcium phosphate crystals are frequently found in osteoarthritic joints (see below).

At present, we have no treatment to influence this process. The knee and hip are often affected and, if the pain and limitation to mobility are severe enough, the orthopaedic surgeon will remove the joint and replace it with an artificial one, usually with very gratifying results.

Inflammatory arthritis

The other major class of articular disease is inflammatory arthritis. This consists of a group of systemic immunological diseases that focus their attention on the synovium of the joint. Rheumatoid arthritis (RA) is the archetypical form of inflammatory arthritis. In RA, there is an imbalance between the proinflammatory and anti-inflammatory cytokines that favours the induction of autoimmunity (an immune reaction against one’s own tissues), chronic inflammation and joint damage. The synovium becomes inflamed, resulting in pain, tenderness, heat and stiffness in the joint as well as a systemic reaction consisting of malaise and fatigue. An acute phase reaction ensues, which can be detected and monitored by measurement of plasma C-reactive protein (CRP) concentration or the erythrocyte sedimentation rate (ESR). In what may well be a separate process, in time the synovial cells increase in size and number, the synovium becomes thickened and develops the changes of chronic inflammation with infiltration by macrophages and lymphocytes. There are also alterations in the small blood vessels. In addition, the amount of synovial fluid in the joint increases in quantity, which adds to the swelling and dysfunction of the affected joint. A number of different cytokines associated with the rheumatoid process can be found in the synovial tissue and fluid; these include tumour necrosis factor alpha (TNF α), interleukins one (IL-1) and six (IL-6). There are also abnormalities of the T and B cell lymphocytes. If left to its own devices, the thickened synovium will spread over the cartilage and erode it and the neighbouring bone. This results in joint damage, deformity and disability.

Non-steroidal anti-inflammatory drugs, e.g. diclofenac, ibuprofen and naproxen are used less frequently than they once were. Although they often provide symptomatic relief, they increase the risk of cardiovascular events, can cause chronic kidney disease and have no influence on the long-term outcome of the disease.

Of more concern in the long term, is the chronic inflammation that results in joint damage and disability. Until about 20 years ago, there was little that we could do to slow or even stop the inexorable progress of this destructive process, meaning that many people with RA became very disabled with joint deformity and pain. However, there have been dramatic improvements in our drug treatment of this most unpleasant condition. We have learnt that we need to diagnose the condition as early as possible. This enables treatment with one or more disease modifying anti-rheumatic drugs (DMARDs) (e.g. methotrexate, leflunomide, sulfasalazine), to be started as soon as possible. As a rule, these drugs rapidly bring the arthritis under control, but different drugs suit different people so sometimes a degree of trial and error is required. DMARDs are taken for long periods and can cause side-effects such as liver damage or bone marrow suppression, and regular blood tests (full blood count, ESR, CRP and liver function tests) are used to monitor their safety as well as the activity of the disease.

More recently, the treatment of RA has been transformed by the introduction of so-called 'biologics' (monoclonal antibodies designed to counteract the effects of cytokines or alter the function of T or B cell lymphocytes). They are expensive and need to be taken long term, so at present they are used in the cases where DMARDs have been unsuccessful. However, in time, they may become the first-line in treatment of RA. Biologics in common use include agents against TNF α (e.g. etanercept) and anti IL-6 (tocilizumab), as well as B cell depleters (rituximab) and T cell blockers (abatacept). This is a field of active research with the potential to transform the treatment not only of RA but of many other immunological diseases.

Other examples of chronic inflammatory arthritis include ankylosing spondylitis, which mainly affects the small joints of the spine, the arthritis associated with psoriasis and the 'reactive' arthritis one sees in response to certain intestinal or genitourinary infections.

The connective tissue diseases

The connective tissue diseases (CTDs) are a fascinating group of conditions in which, for reasons we do not understand, the normal fine balance of the immune system gets disturbed, causing the body to react against itself. In health, the immune system consists of series of counterbalancing mechanisms that marshal defences against alien material like bacteria, viruses or foreign tissue, while not reacting in this way against itself. The identification of foreign tissue (the antigen) is the first step in the process and is followed by a series of defensive reactions including manufacturing antibodies. In CTDs, the person's ability to recognize their own tissue goes awry and they form antibodies against their own tissue (autoantibodies). In some cases, these are responsible for the clinical manifestations of the disease; in others they are a peripheral effect of the underlying disease process. Theories of the pathogenesis of autoimmune diseases are discussed in Chapter 30.

Systemic lupus erythematosus (SLE) mainly affects younger women and can involve one or more of the

bodily systems. It is characterized by the formation of autoantibodies against the cell's nucleus, or part of the nucleus, which are believed to be responsible for many of the manifestations of the disease.

Joint pain is common, although it is unusual to see joint damage such as occurs in RA. SLE can result in a variety of manifestations in other systems, for example various rashes, haemolytic anaemia and glomerulonephritis leading to renal impairment, as well as cerebral, lung and heart problems. Some of these can be life threatening.

Autoantibodies against cell nuclei (antinuclear factor, ANF) are present in the plasma in almost every patient with SLE. However, they are also frequently present in other CTDs and in a proportion of healthy people. Antibodies to specific components of the nucleus are more specific. For example, antibodies against double-stranded DNA are strongly suggestive of SLE.

Management depends on the systems affected. The joint and skin manifestations can often be controlled with hydrochloroquine and/or low doses of prednisolone. More severe cases require the use of high doses of corticosteroids with immunosuppression using cyclophosphamide, azathioprine or mycophenolate mofetil.

Other CTDs are associated with autoantibody formation, although in these cases the disease is not believed to be caused by specific antibody-antigen complexes as in SLE. Examples of these conditions include Sjögren syndrome, polymyositis, scleroderma and primary biliary cirrhosis.

For details and associated immunological abnormalities, see [Table 32.1](#). [Table 32.1](#) is, of necessity, an oversimplified guide to some of the more important signs and symptoms seen in the CTDs, with their associated antibodies. There is often a crossover of antibodies in the various diseases and the antibodies are by no means seen in every case.

Aches and pains

If we move from the laboratory to the clinic or surgery, 'articular disorders' take on another perspective. About one-fifth of the patients in a GP's surgery will complain of musculoskeletal symptoms, such as aches, pain, stiffness or disability. Only a proportion of these patients will have a demonstrable disease or significant anatomical abnormality. Some will have abnormalities associated with the way they move their musculoskeletal system ('mechanical' pain), while in other cases, no abnormality can be identified. This does not mean the symptoms are imagined, but in the absence of an abnormality, attempts to diagnose and 'cure' a disease will, of course, be futile. Doctors are trained to diagnose diseases and then treat them. Too many doctors are uncomfortable when they are not able to make a diagnosis. This may provoke them to order unnecessary investigations in these often anxious patients, thus exposing both parties to the dangers of false positives, laboratory or clerical error and misinterpretation of the results.

As a rheumatologist, each week the author sees one or two patients referred as a consequence of an inappropriately performed blood test. These patients usually have mechanical or unexplained musculoskeletal symptoms

TABLE 32.1 The connective tissue diseases, their common clinical features and frequently associated autoantibodies

Disease	Clinical features	Autoantibodies
SLE	Rashes, joint pains, glomerulonephritis, haemolytic anaemia, leukopenia, neuropsychiatric lupus, pericarditis, pleurisy, vasculitis, fatigue	Anti-nuclear factor Anti-double-stranded DNA Anti-Sm Anti-Ro Anti-La
Scleroderma	Thickened skin, Raynaud phenomenon, pulmonary hypertension, calcinosis, pulmonary fibrosis	Anti-nuclear factor Anti-centromere
Sjögren syndrome	Dry eyes, dry mouth, joint pains, fatigue, Raynaud phenomenon	Anti-nuclear factor Anti-Ro Anti-La
Polymyositis	Muscle weakness and wasting, joint pains, pulmonary fibrosis, fatigue	Rheumatoid factor Anti-nuclear factor Anti-Jo-1 Anti-synthetase
Dermatomyositis	Proximal muscle weakness and wasting, rashes, joint pains, pulmonary fibrosis	Anti-nuclear factor
Wegener granulomatosis	Vasculitis, sinusitis, pulmonary infiltration, glomerulonephritis, rashes	Anti-Mi2 cANCA
Microscopic polyangiitis	Glomerulonephritis, joint/muscle pains, rashes, lung infiltration	pANCA
Primary biliary cirrhosis	Jaundice, itching, joint pains, Raynaud phenomenon	Antimitochondrial

cANCA, cytoplasmic anti-neutrophil cytoplasmic antibody; pANCA, perinuclear anti-neutrophil cytoplasmic antibody; SLE, systemic lupus erythematosus.

and their doctors have performed a batch of tests with vague rheumatological connotations, presumably to reassure themselves or to be seen to be doing something for the patient. The problem is that these tests are often positive in healthy people, especially at low levels. For example, low titres of ANF are widespread, and it is common for a number of anxious young women to be referred with aches and pains, 'weakly positive ANF' and a sheaf of internet misinformation about how fatal SLE is. Informed of their 'positive' test, they have often had to wait several months for their referral appointment, in fear and trepidation, and come prepared to hear the worst. False positive rheumatoid factor tests, high plasma urate concentrations and slightly raised ESRs produce a steady stream of worried well people with fears they have rheumatoid arthritis, gout or inflammatory disease, respectively. Inappropriate tests are not only of concern to the biochemist and accountant. They cause a mass of unnecessary work for the clinician, but the real victims are the unfortunate patients who have to endure long periods of unnecessary strife, stress and uncertainty.

Crystal arthritis

Just as crystals form *in vitro* when the concentration of a solute is sufficiently high to allow nucleation and growth, so do they form *in vivo*. Crystals are sometimes found in joint fluid. These include monosodium urate (MSU), calcium pyrophosphate (CPP), basic calcium phosphates, cholesterol and oxalate crystals. The body's response to crystals varies. In some cases, the presence of crystals in a joint provokes a severe inflammatory reaction, but researchers have found crystals in joint fluid from asymptomatic patients. Different crystals tend to favour different sites. Monosodium urate is preferentially deposited in cartilage and synovium, CPP in articular

fibrocartilage and basic calcium phosphates in tendons and hyaline cartilage. The formation of crystals is influenced by physical factors such as temperature and hydrogen ion concentration (pH), but it is a complex process and what follows is a practical simplification.

Hyperuricaemia and gout

Gout is the best known form of crystal arthritis and is caused by MSU crystal deposition in articular tissues. The term hyperuricaemia is used to describe excessive uric acid or urate in the blood as defined by solubility in plasma at 37 °C; that is, >0.42 mmol/L in males and >0.36 mmol/L in females. While there is a relationship between MSU crystal deposition and hyperuricaemia, it is by no means absolute. Hyperuricaemia without attacks of gouty arthritis is common, and sometimes, particularly during an acute attack, plasma uric acid concentrations are normal.

Uric acid is the end-product of purine metabolism (Fig. 32.2). Purines are components of nucleic acids and of nucleotides that are involved in energy transformation and phosphorylation reactions and act as intracellular messengers. They are derived from diet, the breakdown of nucleotides and from *de novo* synthesis. Purines are metabolized to urate via hypoxanthine and xanthine, the final step being catalysed by xanthine oxidase. The kidneys excrete two-thirds of the urate, the remainder being removed via the bowel, where it is broken down into carbon dioxide and ammonia by bacterial action. The amount of urate pooled in the body depends on the relationship between the input from diet, the breakdown of nucleotides and *de novo* synthesis versus the output via kidney and bowel. Hyperuricaemia results from excess production of urate or reduced excretion, or sometimes both mechanisms. The average uric acid pool size is 7.2 mmol in men

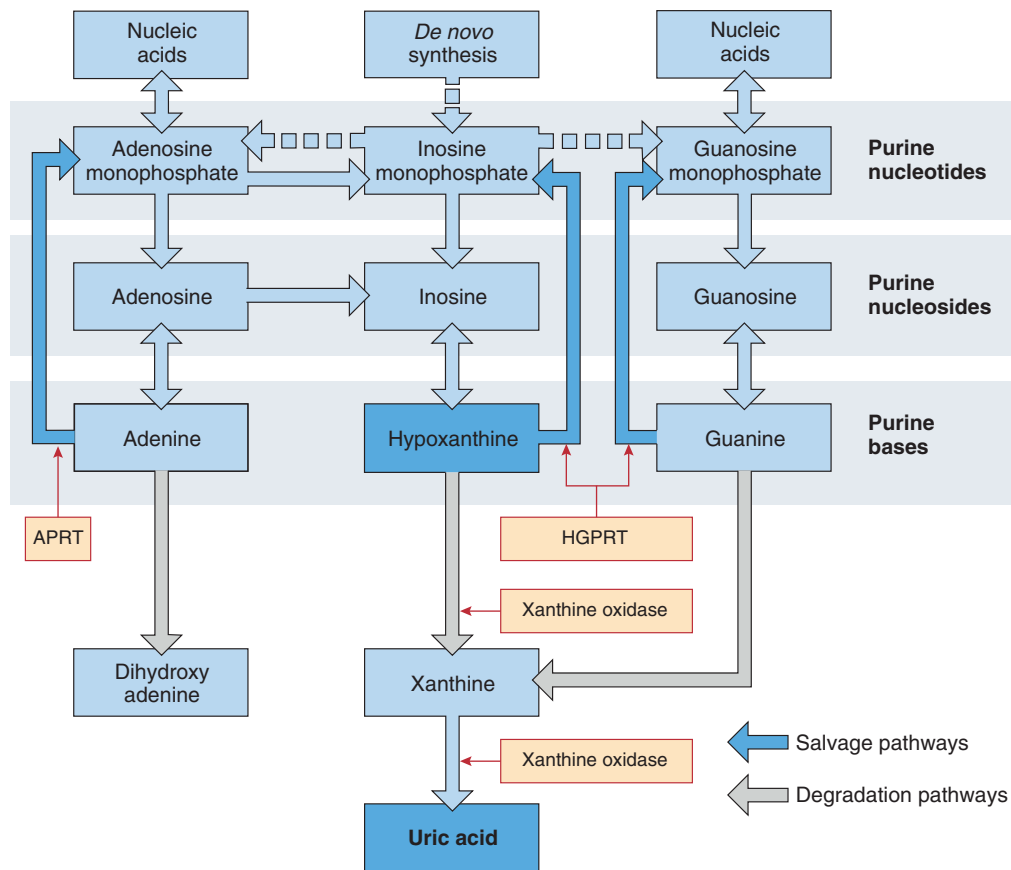


FIGURE 32.2 ■ A simplified diagram of the pathways of purine nucleotide metabolism and uric acid synthesis in humans. APRT, adenine phosphoribosyl transferase; HGPRT, hypoxanthine-guanine phosphoribosyl transferase. (From Marshall W J, Bangert S K, Lapsley M 2008 Clinical Chemistry, 7th ed. Edinburgh: Elsevier, with permission.)

and 3.6 mmol in women. Two-thirds is turned over each day. The uric acid pool of people with gout is increased to 12–24 mmol; those with deposits of urate (tophi) may have a pool as large as 180 mmol.

Gout tends to run in families and is frequent in certain races, such as the New Zealand Maori. This points to an inherited abnormality in urate metabolism, but while there are well-defined enzyme defects resulting in hyperuricaemia, for example hypoxanthine-guanine phosphoribosyl transferase deficiency, the cause of the Lesch–Nyhan syndrome (see p. 640), these are rare. In most cases, we can only divide people with gout into those who undersecrete uric acid from the kidney (85%) and those who synthesize an increased amount.

Gout commonly affects middle-aged males; it is most unusual in premenopausal women. Hyperuricaemia, family history, obesity, hypertension, age, alcohol consumption and renal insufficiency are risk factors for gout. Several drugs, particularly diuretics, low-dose aspirin and ciclosporin, and lead poisoning, may provoke gout by causing hyperuricaemia. Conditions with high purine turnover, such as leukaemia and lymphoma and their treatment with cytotoxic drugs, can cause gout and sometimes acute urate nephropathy by increasing the urate load.

Asymptomatic hyperuricaemia. Asymptomatic hyperuricaemia is the first stage of gout. This is seen on

at least one occasion in as many as 10% of adults, the majority of whom never experience an attack of gout. In the past, we have ignored asymptomatic hyperuricaemia, but it is looking increasingly likely that it is a risk factor for acute coronary syndromes and it is also linked with obesity, type 2 diabetes, hypertension and hyperlipidaemia.

Acute gout. Acute gout usually strikes the first metatarsophalangeal joint (the bunion joint in the big toe) at night. Crystal formation is more likely at colder temperatures and this joint is cooler than more proximal ones of a similar size. The joint rapidly becomes severely painful, swollen, red and exquisitely tender to touch. The weight of the bedclothes is unbearable. There is often a systemic reaction with fever, an acute phase reaction and a leukocytosis. While acute gout most commonly attacks the great toe (70% of cases), the instep, the ankle or knee can be affected. Examination of synovial fluid drawn from an affected joint will show many MSU crystals. Left to itself, acute gout usually subsides over ten days or so.

While some patients never have another attack, there are usually further sporadic attacks. These may become more frequent and may involve several areas at a time (polyarticular gout) as the urate load increases and the disease progresses. In time, upper limb joints such as the fingers, wrists and elbows become affected.

Chronic tophaceous gout. In the later stage of untreated disease, the patient may have continuous inflammation of many joints with associated pain, tenderness and immobility. In addition, crystals of urate are deposited in masses of chalky material called tophi, which sometimes ulcerate and discharge through the skin. These are often seen over joints and bony prominences such as the points of the elbows. They can occur in the joint and inside bone causing permanent damage, pain, deformity and disability. Urate crystals are often the focus for renal stone formation. At this stage, gout is a miserable business.

Diagnosis. When gout presents in the classic manner, the diagnosis is obvious. However, infection in a joint also makes it hot, red and painful and sometimes it can be difficult to differentiate between crystal arthritis and infection. Because both conditions cause fever, leukocytosis and an acute phase reaction, the only reliable way to make the diagnosis is to aspirate fluid from the joint and examine it for crystals (using polarizing light microscopy, see Fig. 32.3) and for bacteria (Gram stain and culture). Rarely, the two conditions coexist, so both tests should be performed.

For the reasons mentioned above, plasma urate concentrations are not helpful in diagnosis, although very high values are convincing and sometimes influence decisions about drug treatment.

Treatment. There are three strands to the treatment of gout. Immediate action is required to counteract the severe inflammation that is producing such pain and disability. In straightforward cases, anti-inflammatory drugs, such as diclofenac or naproxen, can rapidly achieve this. Colchicine, which stabilizes lysosomal membranes, can be used when anti-inflammatory drugs are inadvisable, for example in patients with kidney or stomach conditions. These drugs settle the inflammation but have no effect on crystal formation or the underlying problem of the increased urate load. The second strand to treatment is to

try to persuade the patient to modify his lifestyle. More exercise, less alcohol, improved diet and weight loss are the goals. They are more easily achieved in some than in others.

The third strand of treatment is to reduce the urate load. This will have built up over years and so will take time to reduce. Reduction is usually achieved by the use of allopurinol, a potent purine xanthine oxidase inhibitor, which blocks the conversion of xanthine to urate. Xanthine is soluble and is excreted renally. The decision to start allopurinol should be influenced by the frequency of the attacks of acute gout, the presence of tophi and/or chronic polyarticular gout and very high urate concentrations. Allopurinol gradually leaches away the high urate load and, in time, urate concentrations become normal and tophi are absorbed. There is a danger of allopurinol provoking an acute flare of gout when it is first used, so it should not be started until the acute attack has been controlled. Once treatment with allopurinol has been established, urate concentrations should be checked. The target is a value of <0.3 mmol/L and the dose of allopurinol should be titrated accordingly.

The treatment of uncomplicated gout is simple and rewarding if patients can be persuaded to take their tablets regularly. Male patients often seem to find it difficult to continue their medication when they feel well and this is by far the commonest cause of failure of treatment. However, treating complicated gout can be challenging. These patients often have chronic kidney disease, cardiac disease, peptic ulcers and are taking warfarin, all of which are relative contraindications to the use of anti-inflammatory drugs. They often require diuretics (which may provoke gout) to control heart failure and gout is frequently a problem in transplant patients because of the use of the hyperuricaemia-inducing anti-rejection drug, ciclosporin. In these cases, injecting corticosteroid into the affected joint controls the acute inflammation. If several joints are affected, systemic adrenocorticotrophic hormone (ACTH) or corticosteroids are necessary.

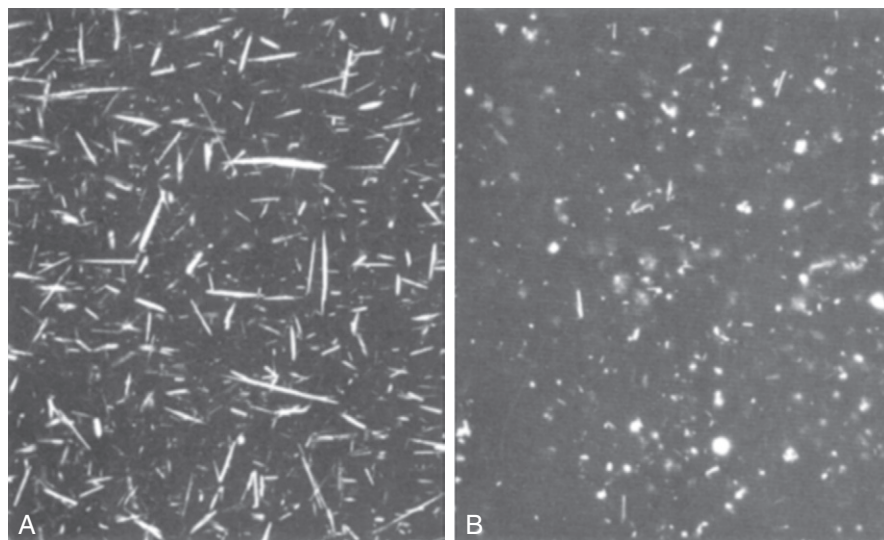


FIGURE 32.3 ■ Sodium urate (needle shaped, A) and calcium pyrophosphate (rhomboid, B) crystals viewed microscopically. Under polarized light, the former are negatively, the latter positively, birefringent.

Lowering urate concentrations with allopurinol has been the mainstay of the treatment of gout for the past 40 years. It is inexpensive and in most cases is very effective. However, drug reactions do occur and there has been a need for other urate lowering agents. Recently, febuxostat, a non-purine selective xanthine oxidase inhibitor, has been introduced and is showing great promise. It does not inhibit enzymes involved in purine and pyrimidine metabolism, as does allopurinol and may in time become the drug of first choice.

All mammals apart from humans and Dalmatian dogs possess the enzyme urate oxidase (uricase) which oxidizes urate to allantoin. This is 5–10 times more soluble than uric acid and so is more effectively excreted by the kidney. However, urate oxidase was lost in early human evolution. The reasons for this are not clear but it has been proposed that its loss was an advantage because uric acid is a powerful antioxidant and scavenger of free radicals that might protect against oxidative damage and so prolong life and decrease the risks of cancer. As a result, humans are prone to gout. The selective inbreeding of Dalmatian dogs has produced a striking white animal with black spots but also one without urate oxidase. Hence, the Dalmatian's tendency to the development of gout and the formation of urate renal stones, which can be treated with allopurinol.

Until recently, this interesting information was of peripheral interest. However, researchers have investigated the use of urate oxidase to treat hyperuricaemia in humans. There are two preparations. Rasburicase, a recombinant form from the fungus *Aspergillus flavus*, is used to prevent acute urate nephropathy, which can occur when large tumour loads are treated with chemotherapy (tumour lysis syndrome, see p. 114). However, its very short circulating life and the fact patients sometimes develop antibodies to the drug limits its use. These disadvantages led to the development of PEG-uricase, a recombinant porcine urate oxidase, to which multiple strands of polyethylene glycol (PEG) of average molecular weight 10 000 Da have been attached. This greatly prolongs its circulating life and it is non-immunogenic. However, it has to be given intravenously at two-weekly intervals and can cause flares of gout and infusion reactions. At present, its use is reserved for patients with gout that is difficult to control who cannot take standard treatment. It may be noted that accurate measurement of urate in patients treated with this drug is difficult, as it continues to act *in vitro*.

Calcium pyrophosphate deposition (CPPD)

This is the other common form of crystal arthritis. It is not as frequent as gout and is less well understood. Calcium pyrophosphate crystals are preferentially deposited in articular cartilage where calcification can be seen on X-rays (chondrocalcinosis). Affected joints can become acutely inflamed in a way that resembles gout (hence CPPD's other name of pseudogout). The knee and the wrist are most frequently affected. An acute phase reaction is common and, again, finding CPP crystals in the joint fluid makes the diagnosis. In some cases, there is chronic involvement of several joints, which can cause diagnostic confusion with other inflammatory articular diseases.

Calcium pyrophosphate deposition becomes more common with age. Symptomless chondrocalcinosis is often seen on X-rays of the elderly. Attacks are sometimes triggered by trauma unrelated to the joint (of which surgery is an excellent example), and rheumatologists are often asked to see elderly postoperative patients in the surgical wards with a hot, swollen, 'infected' wrist or knee. Almost invariably, pseudogout is responsible.

Familial CPPD is well recognized, and kindred from many different countries have been described. These patients present in their twenties and usually have severe destructive joint disease. In some families, a genetic abnormality on the short arm of chromosome 5p has been identified.

Calcium pyrophosphate deposition is associated with several metabolic diseases described in other chapters. Definite associations have been established with haemochromatosis, hyperparathyroidism, hypomagnesaemia, hypophosphatasia, alkaptonuria and Wilson disease. In these cases, the onset is often at a younger age (under 55 years) and the disease is more severe than is seen in uncomplicated cases. A link with thyroid disease is now disputed.

As yet, we have little understanding of the underlying biochemical processes at play in this group of conditions. Our 'first aid' style of treatment, with anti-inflammatory drugs and corticosteroid injection of affected joints, reflects this.

Basic calcium phosphate deposition disease

Basic calcium phosphate (BCP) crystals include hydroxyapatite, octacalcium phosphate and tricalcium phosphate. They are often difficult to identify, which is why we still have so much to learn about these conditions. Basic calcium phosphate crystals are sometimes deposited in the tendons or in soft tissues outside the joints, where they can give rise to intense pain. This tends to occur in younger females and affects the shoulder, wrist and even the great toe, when gout may be incorrectly diagnosed. Injection of corticosteroid into the lesion often results in dramatic improvement. Mercifully, the pain usually settles down over a week or two and the calcification seen on X-ray is often resorbed within six weeks. Basic calcium phosphate crystal deposition has also been implicated in an unusual condition that causes large effusions and joint resorption in the shoulders of elderly women (Milwaukee shoulder).

There is increasing interest in these crystals because they appear to be closely involved in the process of osteoarthritis (OA). Basic calcium phosphate crystals can be found in the synovial fluid of as many as 60% of joints severely affected by OA. It is uncertain whether this is related to the cause of OA or is a downstream effect of the OA process. However, in tissue culture, BCP crystals stimulate mitogenesis and the production of metalloproteinases, which cleave cartilage matrix. This suggests BCP crystals may be involved in the cartilage breakdown seen in OA, and researchers are starting to look at the effects of agents such as phosphocitrate, which inhibit the formation of BCP crystals.

Other crystals found in synovial fluid

Other crystals can be found in synovial fluid. These include cholesterol and oxalic acid crystals. These latter occur in primary oxalosis and in patients with established renal failure on dialysis, in whom the crystals can provoke a chronic polyarthritis in the small joints of the hand.

ARTICULAR INVOLVEMENT IN ENDOCRINE AND METABOLIC DISEASES

Diabetes mellitus

Over 30% of diabetic patients have potentially disabling hand or shoulder disorders. Capsulitis of the shoulder (frozen shoulder) is particularly common in diabetes. It causes severe, disabling shoulder pain, which settles after a few months but leaves the patient with very limited mobility at the shoulder joint. The range of movement usually improves over the months. Sometimes a corticosteroid injection will help the pain, but limited movement remains a problem.

Many patients with diabetes develop changes in the connective tissues of the hands. Dupuytren contracture, a condition in which there is thickening of the palmar fascia and increasing contracture of fingers, is common in diabetes. In what seems to be a different process, called diabetic cheiroarthropathy or diabetic hand syndrome, the skin and underlying tissues become thickened, resulting in a limited range of movement of the hand and finger joints. Often the finger tendons are affected. These changes are sometimes mistaken for rheumatoid arthritis but there is no joint damage or acute phase reaction. Biopsy shows excessive fibrosis and increased deposition of dermal collagen. It has been suggested that increased tissue glycation may result in diminished collagen breakdown. Perhaps a similar sort of process is taking place in the frozen shoulders of diabetic patients.

Peripheral nerve damage is a frequent complication of diabetes. In some cases, the nerve supply to the joints is affected so that the usual sensations of joint position and joint movement are lost. Repeated trauma can take place within such a joint without the patient being aware and this can rapidly result in painless destruction and bony resorption of large areas of bone around the affected joint. This is called a Charcot joint. It occurs almost exclusively in the foot and is a major problem. However, although evidence of damage to small nerve fibres is usually present in this condition, other factors are undoubtedly involved. Charcot arthritis occurs more frequently in type 1 diabetes. The reason for this is not clear.

Other endocrine disorders

Several endocrine disorders can cause muscle and joint pains and need to be considered in the assessment of the patient with aches and pains. Only rarely will they be found to be responsible, but early diagnosis and treatment will often relieve the patient's symptoms and can

prevent further suffering and tissue damage. As ever with uncommon conditions presenting in unusual ways, the important thing is to think of the diagnosis.

Both hypo- and hyperthyroidism can result in joint pains and a variety of muscular symptoms. Hyperparathyroidism is a cause of bone pain and aching joints as well as chondrocalcinosis and pseudogout, while joint pains, degenerative arthritis, joint laxity and muscle weakness can result from effects of acromegaly on bone, joints and soft tissue. Hypercortisolism sometimes presents in the rheumatology clinic with back pain (from osteoporotic fractures) and muscle weakness.

Haemochromatosis

As described in Chapter 14, liver disease and diabetes in a male with slate grey skin is the classic presentation of haemochromatosis. However, the alert rheumatologist will discover the occasional patient with this condition presenting in the clinic with articular disease. This can happen in two ways. We have already discussed how haemochromatosis is associated with chondrocalcinosis and pseudogout. In addition, haemochromatosis itself is responsible for a characteristic form of arthritis involving the metacarpophalangeal joints of the index and middle fingers. X-rays show loss of joint space and characteristic hook osteophyte formation in these joints. There is no inflammatory reaction, but iron studies show the characteristic picture of haemochromatosis. The synovium contains much iron. Venesection unfortunately makes little difference to the joint symptoms, but does prevent further harm to the pancreas and liver. It is also important to screen the patient's relatives so that those who have inherited the disease can be identified and treated before damage occurs.

Alkaptonuria

This rare, autosomal recessive disorder is caused by a deficiency of the enzyme homogentisic acid oxidase. This results in a build-up of homogentisic acid, an intermediary product in the metabolism of phenylalanine and tyrosine. The excess homogentisic acid is oxidized and polymerized and forms a blackish pigment, alkapton, which is deposited in cartilage and can be easily seen in the cartilage of the ears. More importantly, it is also deposited in articular cartilage, which becomes more susceptible to mechanical stress and degeneration. Arthritis of the hips, knees and shoulders develops early. Alkapton is also deposited in the intervertebral discs in the spine, which become calcified. This produces a characteristic X-ray appearance and predisposes to mechanical back problems. The diagnosis is made by identifying homogentisic acid in the urine.

LABORATORY TESTING IN ARTICULAR DISEASE

Anaemia in rheumatoid arthritis

Anaemia is common in RA. When the disease is active, there is frequently a normochromic, normocytic anaemia. This is called 'the anaemia of chronic disease'

(see Chapter 27) and it will resolve if the activity of the arthritis can be controlled. Iron is present in the bone marrow but cannot be utilized in erythropoiesis.

Some RA patients require an anti-inflammatory drug to keep their pain and stiffness at bay. These drugs have a tendency to cause erosions and ulcers in the stomach, duodenum or small bowel, resulting in blood loss, iron deficiency and a hypochromic microcytic anaemia.

Separating iron-deficiency anaemia and the anaemia of chronic disease in these patients is often tricky and is made more difficult by the behaviour of ferritin in inflammatory diseases. Ferritin is one of the several proteins that are acute phase reactants, and for this reason, its plasma concentration is usually raised when RA is active. Measurement of plasma ferritin concentration is considered to be the best test of iron deficiency (see Chapter 27), but in RA, the result should be interpreted with caution. Usually a concentration of $<12 \mu\text{g/L}$ is taken as diagnostic of iron deficiency, but in RA, rheumatologists take a concentration of $<30 \mu\text{g/L}$ to indicate iron deficiency. However, iron deficiency can exist with higher concentrations of ferritin. The presence of microcytosis favours the diagnosis of iron deficiency, but sometimes examination of bone marrow for iron stores is the only reliable way to differentiate between iron-deficiency anaemia and the anaemia of chronic disease.

The acute phase response

The acute phase response is discussed in detail in Chapter 30. Traditionally, rheumatologists have used the ESR to diagnose and then monitor an acute phase response. This is, of course, an indirect measure of a series of acute phase proteins and immunoglobulins and is particularly influenced by plasma fibrinogen concentrations. Measurement of CRP concentration is gradually superseding the ESR and is believed to correlate better with disease activity in RA. In SLE, CRP concentration is often normal; indeed, a raised value in this condition is said to suggest infection rather than increased disease activity. In truth, most rheumatologists like to use both CRP and ESR.

These simple investigations are of great help to a rheumatologist. One of our main tasks when confronted with a new patient is to differentiate between the inflammatory and non-inflammatory types of arthritis. This is not always easy and the CRP and/or ESR often provide the answer. The activity of rheumatoid arthritis waxes and wanes over time and this is reflected in the acute phase response, as is the success or otherwise of drug treatment.

Examination of synovial fluid

An acutely swollen hot joint can be a result of infection, crystals or the irritative effects of blood on the synovium. The only reliable way to determine which is responsible is to aspirate the joint fluid. Blood will be apparent to the naked eye. Gram-staining will show bacteria and polarizing light microscopy will demonstrate crystals (Fig. 32.3) and whether they are negatively (gout) or positively (CPP) birefringent. The chances of identifying crystals are increased if the specimen is centrifuged and the deposit examined at the earliest opportunity. Sometimes it

is only possible to obtain a tiny amount of fluid. If expressed on to a microscope slide at the bedside and covered with a cover slip to avoid evaporation, this can yield vital information when examined in the laboratory.

Rheumatoid factor

Rheumatoid factors (RFs) are antibodies directed against the Fc portion of human IgG and are present in 75–80% of patients with RA. At present, it is believed that RF is produced as a downstream product of the rheumatoid process rather than being central to the disease process. Rheumatoid factors also occur in other connective tissue diseases, chronic infections such as tuberculosis and hepatitis C and conditions associated with hyperglobulinaemia, such as hyperglobulinaemic purpura, sarcoidosis and cryoglobulinaemia. They are found in low titres in 5% of the population, occurring more frequently with advancing age.

A positive RF is strong confirmatory evidence of RA. In a patient with polyarthritis it has a sensitivity of 70% and a specificity of 80%. However, in an unselected population its predictive value is low (20%). Very high titres of RF are associated with severe disease and are a marker of poor prognosis.

Most laboratory methods identify IgM antibodies, although IgA and IgG antibodies of uncertain clinical significance do exist. For many years, rheumatoid factor was detected by techniques using flocculation or agglutination of IgG-coated particles or cells. The Rose Waaler test, the standard method for many years, used sheep red blood cells coated with rabbit IgG, which agglutinate in the presence of RF, while the flocculation of latex beads coated with human IgG forms the basis of the RA latex test. The sheep cell tests are more specific. However, heterophile antibodies reacting directly against sheep cells sometimes give rise to false positives. Because of this, sera are usually tested on uncoated and coated sheep red cells in parallel. Laser nephelometry, enzyme linked immunosorbent assay (ELISA), immunofluorescence and radioimmunoassay can also be used to measure rheumatoid factors.

Other autoantibodies are seen in RA. In recent years, it has become standard practice to measure antibodies to cyclic citrullinated protein (anti-CCP) as well as RF. These are measured by peptide-based ELISA. The test has a sensitivity of 74% and a specificity of 94%. It may eventually supersede the time-honoured RF test but most rheumatologists use both tests at present.

Other autoantibody tests

Autoantibodies are produced in the connective tissue diseases (see above) and some cases of vasculitis. These are sometimes diagnostic, but most autoantibodies occur in several of these conditions and in healthy people. These tests are valuable in confirming a suspected diagnosis and negative tests are useful in helping to rule out diagnoses. However, because they occur in a proportion of normal people, they should not be used for screening purposes or in cases without clinical features suspicious of connective tissue disease (see p. 639). For some years, autoantibodies have been identified by indirect immunofluorescence

using Hep-2 cells. In this age of automation, there is a move to streamline the process of identifying anti-nuclear antibodies. In some laboratories, automated multiplex immunoassay is used to screen for a mixture of common nuclear antigens. Positive specimens are then examined by direct immunofluorescence, which remains the gold standard. At this stage, the presence of anti-nuclear factor (ANF) is a non-specific indicator that a connective tissue disease may be present (see Table 32.1). The next step is to determine to which part of the nucleus the antibody reacts, by using highly specific tests, including enzyme linked immunoabsorbent assay (ELISA) and indirect immunofluorescence. Anti-double stranded DNA, anti-Sm, anti-Ro, anti-La, anti-centromere, anti-Jo1, anti-Mi2 and anti-mitochondrial antibodies can be identified in this way, each of which, as shown in Table 32.1, is associated with different connective tissue diseases.

CONCLUSION

In spite of the fact that biochemical abnormalities themselves are responsible for few articular diseases, we have seen how clinical biochemistry plays an important part in the diagnosis, treatment and monitoring of rheumatological conditions. Many of the chapters in this book refer to medical specialties that primarily focus on biochemical disorders, or conditions with major biochemical consequences. However, there are some, such as rheumatology, for which this is not the case. It is instructive to examine the different ways in which biochemical tests are useful in such a setting, using articular disorders as an example:

1. Confirming a diagnosis:
 - finding urate crystals in synovial fluid makes the diagnosis of gout
 - finding raised iron and ferritin concentrations in haemochromatosis arthritis and demonstrating the presence of increased iron on liver biopsy confirms the diagnosis
 - finding homogentisic acid in the urine confirms the diagnosis of alkaptonuria.
2. Categorizing class of arthritis:
 - in a patient with arthritis or possible CTD, demonstrating an acute phase response by raised CRP and/or ESR supports the diagnosis of an inflammatory disease, while a normal CRP/ESR favours a diagnosis of osteoarthritis, a mechanical disorder or 'unexplained symptoms'
 - the presence of ANF suggests one of the connective tissue diseases.
3. Supporting a diagnosis:
 - when a diagnosis is suspected, a positive sensitive test is strong support for that diagnosis, for example a positive rheumatoid factor and/or CCP in suspected RA or a raised urate in gout.
4. Excluding a diagnosis:
 - normal iron studies virtually exclude the diagnosis of haemochromatosis.

5. Monitoring the activity of a disease:
 - the activity of inflammatory arthritis waxes and wanes spontaneously. This can be monitored by regular CRP/ESR measurements.
6. Monitoring the effect of treatment:
 - similarly, the CRP/ESR can be used to monitor the effectiveness of drug treatment in controlling the activity of inflammatory arthritis.
7. Monitoring for drug toxicity:
 - regular blood testing is used to monitor for drug toxicity. For example, two-monthly liver function tests are performed routinely on the many patients with RA, who are taking the potentially hepatotoxic drug methotrexate.
8. Monitoring for the complications of a disease:
 - glomerulonephritis and renal failure are fairly frequent complications of SLE. Hence regular tests of renal function and urine tests are part of the routine follow-up of these patients.
9. Specialty-specific idiosyncrasies:
 - each specialty or disease has its own idiosyncrasies, for example the conflicting effect of iron deficiency and the acute phase response on ferritin concentrations in patients with RA and anaemia.

These examples show how powerful biochemical investigations can be, when performed for a defined reason. However, we have also demonstrated that, when used inappropriately, they have the power to inflict harm and distress on the unfortunate victim. We must recall Hippocrates' dictum 'primum non nocere' (first do not harm) at all times.

Further reading

- Dalbeth N, Haskard DO. Pathophysiology of crystal-induced arthritis. In: Wortmann RL, Schumacher Jr HR, Becker MA et al. editors. Crystal-induced arthropathies: gout, pseudogout, and apatite-associated syndromes. New York: Taylor & Francis; 2006.
- Gives in-depth details of the pathophysiology of crystal-induced arthritis.*
- Maddison PJ, Huey P. Serological profile. In: Isenberg DA, Maddison PJ, Woo P et al. editors. Oxford textbook of rheumatology. 3rd ed. Oxford: Oxford University Press; 2004.
- Although written some time ago, this well crafted chapter still provides an excellent and pertinent review of autoantibody testing.*
- Shipley M. Hyperuricaemia and gout. J R Coll Physicians Edinb 2011;41:229-33.
- An up-to-date review of topic.*
- The following three articles are pertinent to the debate about the utility of traditional immunofluorescence microscopy compared with the newer automated methods. While automation is quicker and less labour-intensive, in the present state of our knowledge, immunofluorescence remains the gold standard.*
- Bonilla E, Francis L, Allam F et al. Immunofluorescence microscopy is superior to fluorescent beads for detection of antinuclear antibody reactivity in systemic lupus erythematosus patients. Clin Immunol 2007;124:18-21.
- Kumar A, Bhatia A, Walker Minz R. Antinuclear antibodies and their detection methods in diagnosis of connective tissue diseases: a journey revisited. Diagn Pathol 2009;4:1.
- This article provides an overview on advancement in antinuclear antibody detection methods, their future prospects, advantages, disadvantages and guidelines for use of these tests.*
- Wallace DJ. New methods for antinuclear antibody testing: does it cut costs and corners without jeopardizing clinical reliability? J Immunol Methods 2006;311:189-97.

Muscle disease

Laurence A. Bindoff

CHAPTER OUTLINE

INTRODUCTION 646

FUNCTIONAL ANATOMY AND PHYSIOLOGY OF MUSCLE 646

DISEASES OF MUSCLE AND THEIR INVESTIGATION 650

BIOCHEMICAL INVESTIGATION OF MUSCLE DISEASE 651

'Routine' biochemical studies 651

Plasma creatine kinase activity 651

Other enzymes measurable in plasma 653

Myoglobinuria 653

INVESTIGATION OF MUSCLE DISEASE 653

Non-metabolic, genetically determined myopathies 653

Metabolic, genetically determined myopathies 654

CONCLUSION 659

APPENDIX 659

INTRODUCTION

Diseases affecting striated muscle are important causes of morbidity and mortality. They are common in clinical practice, and patients may be seen by a variety of clinical specialists, including neurologists, rheumatologists, orthopaedic surgeons and paediatricians. Investigating suspected muscle disease requires a combination of clinical and laboratory skills, including biochemical, genetic and pathological investigations, each assuming a different importance depending on the nature of the disorder. In some, biochemical studies play a minor role whereas in others, particularly the metabolic myopathies, biochemical investigations are crucial.

FUNCTIONAL ANATOMY AND PHYSIOLOGY OF MUSCLE

Skeletal muscle accounts for approximately 40% of total body weight and between 30% and 40% of total body oxygen consumption, even at rest. It is, therefore, an extremely important tissue in metabolic terms. Muscle is composed of multinucleated fibres that contain the contractile apparatus upon which movement depends. Although similar in structure, muscle fibres vary and three main types have been defined using metabolic and functional criteria (Fig. 33.1 and Table 33.1). Most skeletal muscles contain all three fibre types, although the proportions vary considerably depending on the function of the particular muscle.

The main function of muscle is to generate force in a controlled manner. This force, in the form of contraction (shortening), is produced in muscle fibres by the

interaction of actin and myosin (Fig. 33.2), a process that is highly energy dependent. The energy required for muscle contraction comes from the hydrolysis of ATP, and maintenance of ATP concentration is critical. Any interference with ATP generation will inevitably impair the ability of muscle to produce force. ATP concentration is maintained by one of two mechanisms: ATP can be regenerated either from the energy storage molecule phosphocreatine or from ADP, or produced directly during glycolysis and mitochondrial oxidation. Regeneration is rapid, while the second process takes longer.

Phosphocreatine is present in large quantities in muscle (the other major site is brain) and acts as a reservoir of

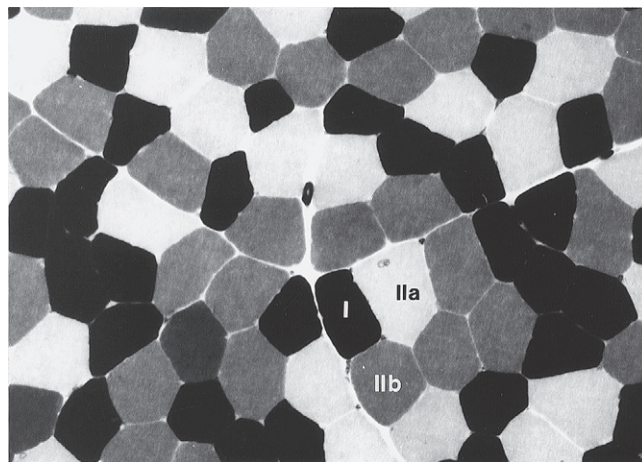


FIGURE 33.1 ■ Cross-section of normal muscle showing the histochemical reaction for ATPase activity. The three fibre types can be easily identified: type I (dark), type IIa and type IIb. (Photograph courtesy of Dr M A Johnson.)

TABLE 33.1 Muscle fibre types

Fibre type	Electrical properties	Metabolic properties	Histochemical properties
Type I	Slow twitch	Oxidative	Heavily stained ATPase after preincubation at acid pH (e.g. pH4.6)
Type IIA	Fast twitch	Oxidative/glycolytic	Lightly stained ATPase after preincubation at pH4.6
Type IIB	Fast twitch	Glycolytic	Lightly stained ATPase only after preincubation at pH4.3

The three main fibre types are differentiated on muscle biopsy by the calcium-activated myosin ATPase. Type I fibres are dependent upon oxidative metabolism, contain the most mitochondria and are particularly important for endurance exercise. Type II fibres contain a higher proportion of glycolytic enzymes and are important in rapid movement. Regenerating fibres have an intermediate staining pattern and are referred to as type IIC fibres.

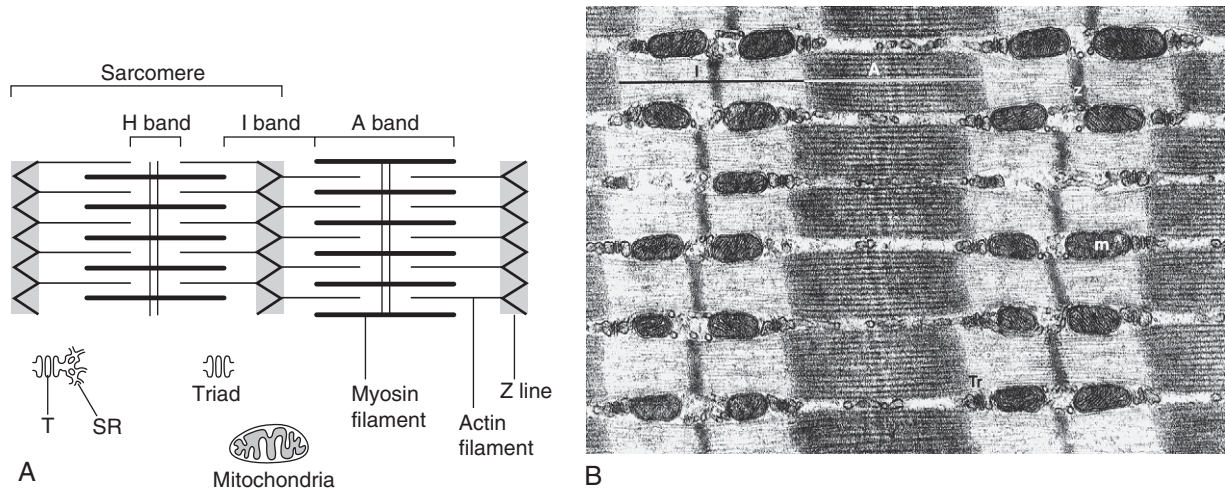
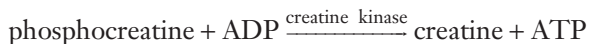
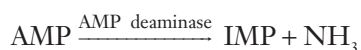
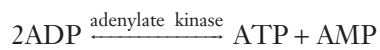


FIGURE 33.2 ■ The structure of skeletal muscle. (A) Diagrammatic representation of the structural components within a single muscle fibre. Thin filaments (actin) are anchored to the Z line. Thick filaments are composed of multiple myosin molecules, each of which has a hinged end that can interact with the thin filament. The T tubule is a continuation of the sarcolemmal membrane, which interacts with the sarcoplasmic reticulum at specific sites (triad). The signal for contraction is transmitted along the sarcolemma to the T tubule, which in turn causes release of calcium from the sarcoplasmic reticulum (SR). Calcium release stimulates contraction. (B) Electron micrograph of normal muscle. This shows the bundles of myofilaments with Z lines (z), thin (I band) and thick (A band) filaments. Triad (Tr) and mitochondria (m) are identified. (Photograph courtesy of Dr M J Cullen. Magnification $\times 30\,000$.)

high-energy phosphate groups that it can donate to ADP in the following transphosphorylation reaction, catalysed by the enzyme creatine kinase:



The concentration of ATP does not fall significantly until nearly all of the phosphocreatine has been converted to creatine. A second transphosphorylation reaction, catalysed by adenylate kinase, seems to have a minor role in ATP production. The AMP formed is broken down further by AMP deaminase.



ATP can also be generated directly by glycolysis and the oxidative catabolism of carbohydrate and lipid fuels. These processes are slow compared with transphosphorylation, but nevertheless are essential for ATP generation. The breakdown of carbohydrate by glycolysis (Fig. 33.3) has a vital role in skeletal muscle since it permits ATP

production under anaerobic conditions. When oxidative metabolism of pyruvate is impaired, for example during ischaemia, which includes high intensity activity, or in the presence of a defect of the respiratory chain, increasing amounts of lactate may be produced.

While small amounts of ATP can be generated by glycolysis in the cytosol, significantly greater amounts are produced by the oxidative breakdown of metabolic fuels (pyruvate, ketone bodies, fatty acids) that occurs within mitochondria. Long chain fatty acids, either from intracellular lipid stores or imported from the bloodstream, are first activated to their acyl-CoA esters before being transported into the mitochondrial matrix by the concerted action of carnitine palmitoyltransferase I, carnitine/acylcarnitine translocase and carnitine palmitoyltransferase II (Fig. 33.4). Short and medium chain fatty acids enter the mitochondria as the free acids and are activated to their acyl-CoA esters in the mitochondrial matrix. Inside the mitochondria, fatty acyl-CoA esters undergo β -oxidation, a series of four reactions that results in the production of acetyl-CoA and a chain-shortened fatty acid (Fig. 33.5); there are two or three enzymes with overlapping substrate specificities for each of these steps. Reducing equivalents

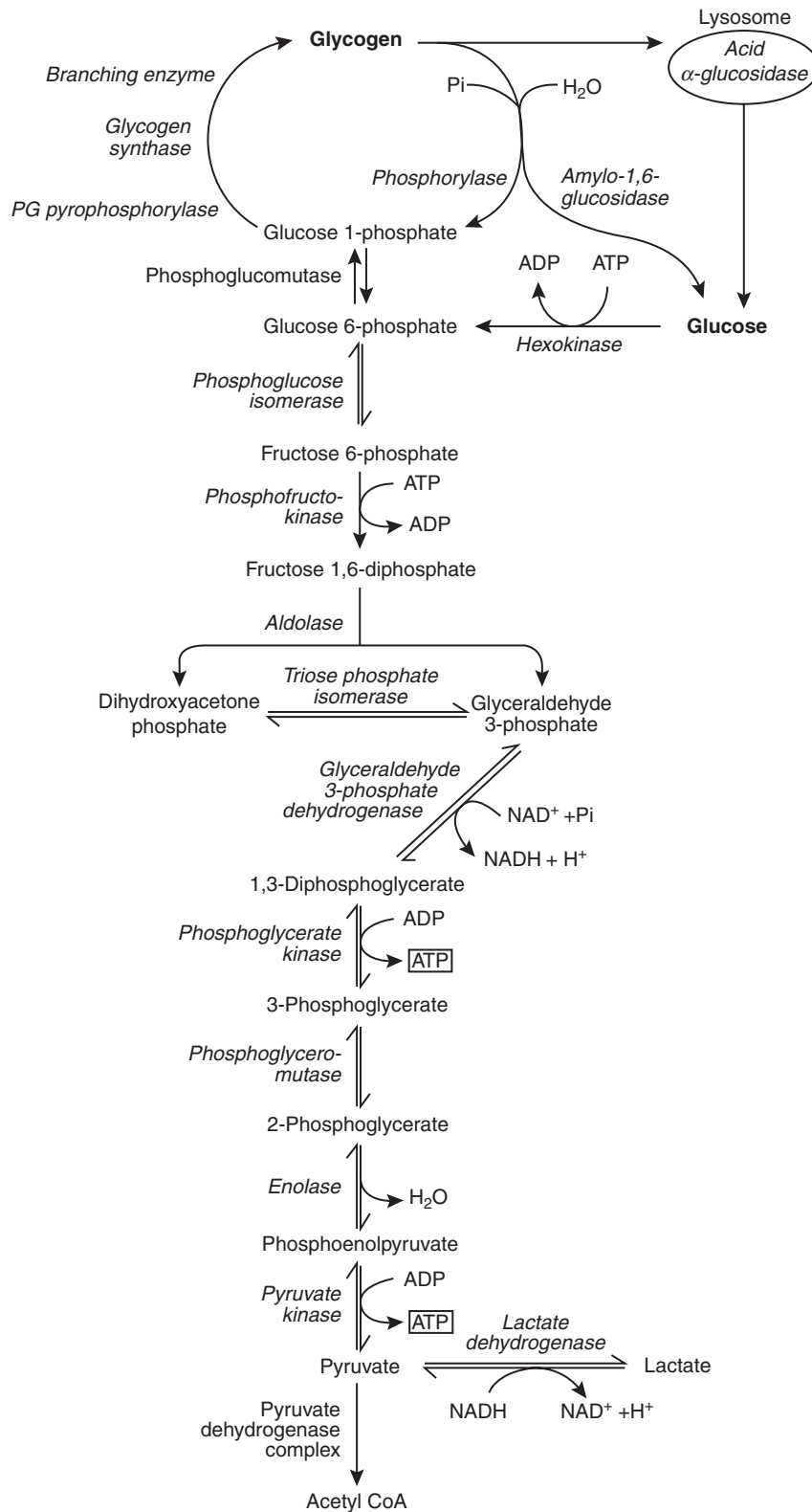


FIGURE 33.3 ■ Glycogen breakdown and glycolysis. Under anaerobic conditions, when further oxidation of acetyl-CoA is impaired, pyruvate is metabolized to lactate. This will also occur in defects of pyruvate dehydrogenase complex (PDC) and of the respiratory chain, and is due to the failure to reoxidize intramitochondrial NADH, which in turn inhibits PDC. NADH formed during glycolysis (and other substrate oxidation) is reoxidized by complex I of the mitochondrial respiratory chain (see Fig. 33.7).

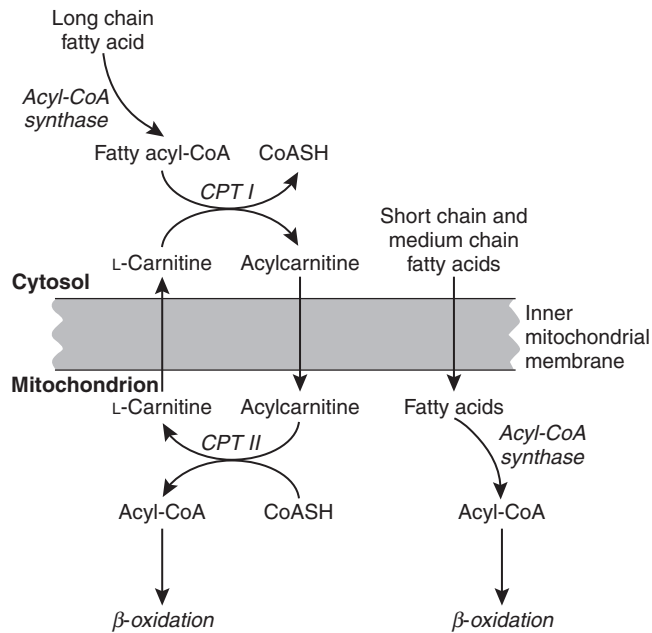


FIGURE 33.4 ■ Transport of fatty acids across the inner mitochondrial membrane. Short and medium chain fatty acids do not require a specialized transport mechanism to cross the membrane. Long chain fatty acids are first acylated in the cytosol. Carnitine palmitoyltransferase (CPT I) on the outer side of the inner membrane converts the fatty acyl-CoA to an acylcarnitine ester and this is transported across the membrane linked to the export of carnitine. Inside the matrix, CPT II converts the acylcarnitine back to the fatty acyl-CoA, which is broken down by β -oxidation.

generated by the process of β -oxidation are transferred to the respiratory chain. The acyl-CoA dehydrogenases transfer reducing equivalents to electron transfer flavoprotein (ETF) and thereafter to ETF dehydrogenase, which directly reduces ubiquinone of the respiratory chain (see Fig. 33.7). The 3-hydroxyacyl-CoA dehydrogenases reduce NAD^+ to give NADH, which transfers its reducing equivalents to complex I of the respiratory chain. Acetyl-CoA generated either from carbohydrate or fatty acid oxidation is metabolized further by the tricarboxylic acid cycle (Fig. 33.6). The oxidation of fatty acids and glucose, as well as the subsequent metabolism of acetyl-CoA, generates more reduced cofactors (NADH and FADH_2) that are re-oxidized by the respiratory chain, and the energy released by this process is conserved as ATP (Fig. 33.7).

The balance of muscle metabolism depends on the state of activity, diet and the influence of various hormones (particularly insulin, thyroxine, glucocorticoids). At rest, muscle predominantly oxidizes fatty acids to generate the energy for ATP synthesis. During exercise, the proportion of energy derived from carbohydrate or lipid depends on the degree and duration of this exercise and on the degree of physical fitness. High-intensity exercise at close to maximum oxygen uptake relies almost exclusively on carbohydrate metabolism, and glycogen depletion coincides with exhaustion. During moderate-intensity exercise for prolonged periods, there is a switch from carbohydrate to lipid metabolism.

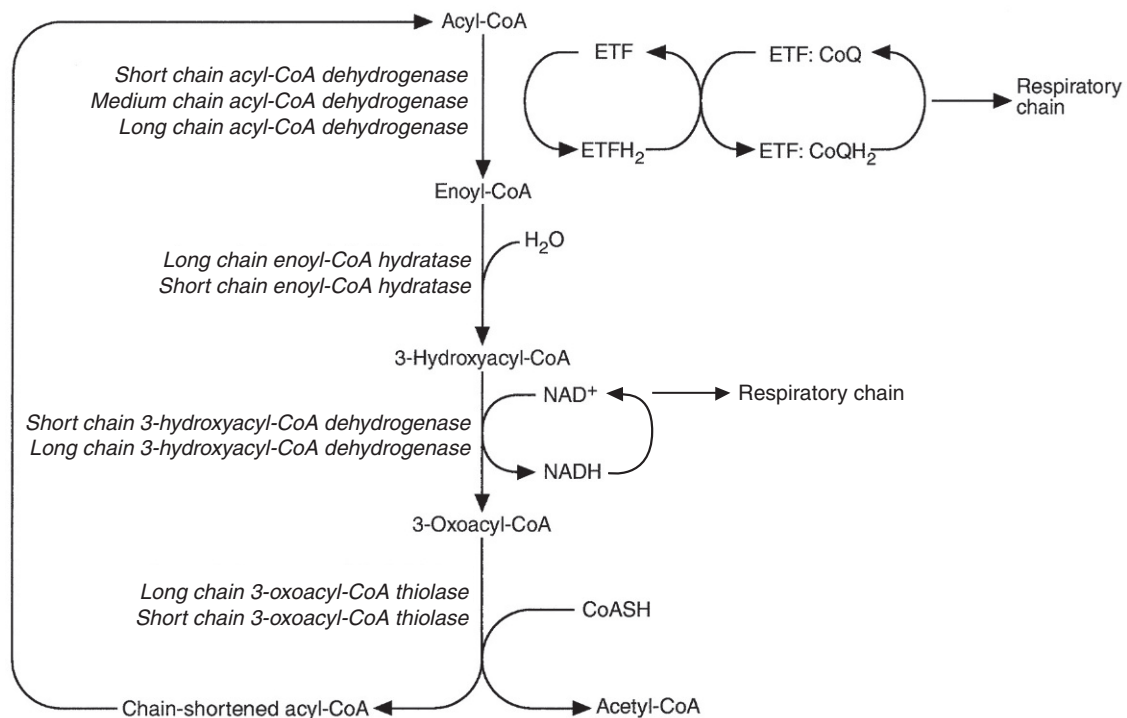


FIGURE 33.5 ■ Mitochondrial β -oxidation of saturated fatty acids. ETF, electron transfer flavoprotein; CoQ, coenzyme Q.

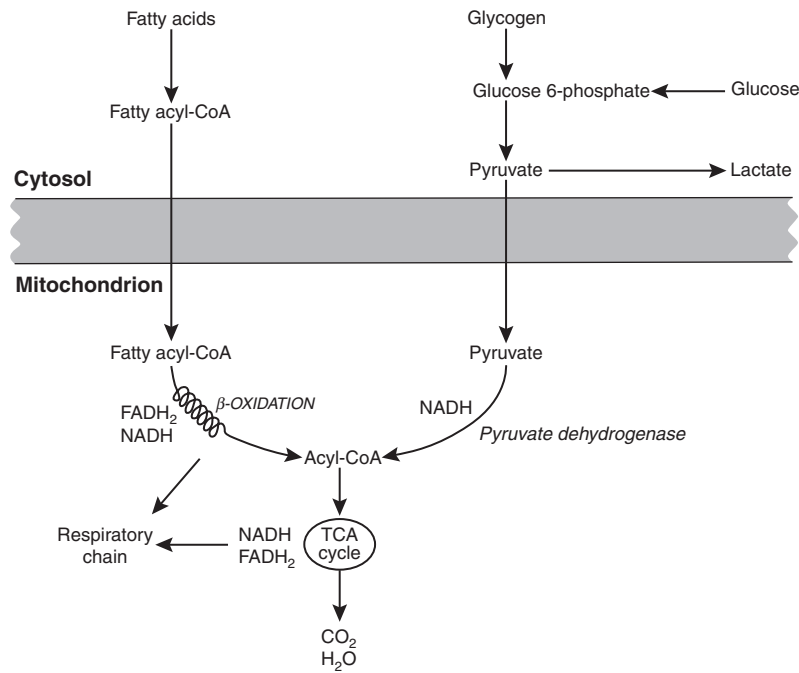


FIGURE 33.6 ■ The production and further metabolism of acetyl-CoA.

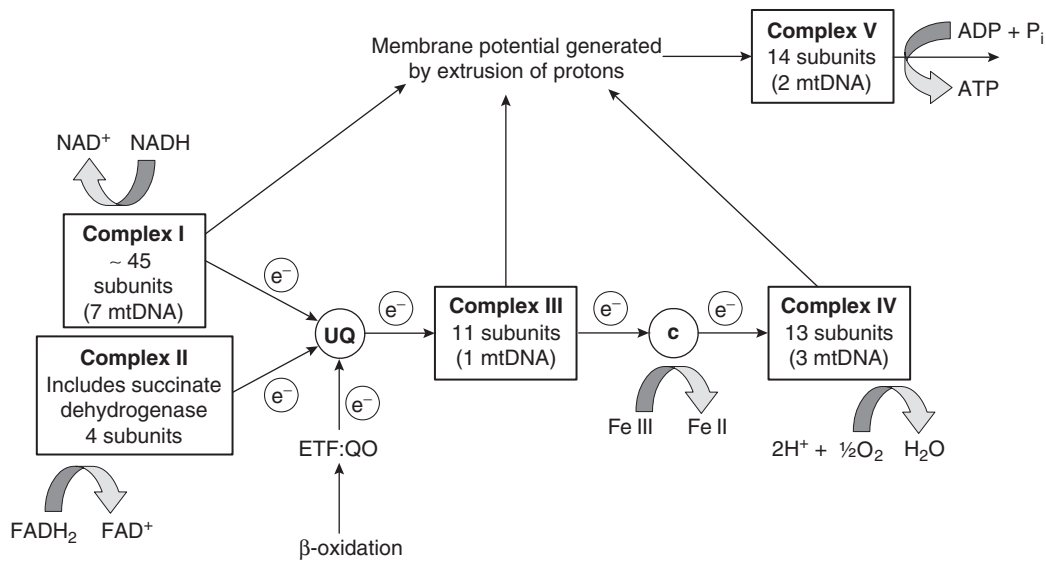


FIGURE 33.7 ■ Components of the mitochondrial respiratory chain. NADH is reoxidized by complex I, while FADH₂ donates electrons via complex II and ETF dehydrogenase (ETF:QO) donates electrons directly to ubiquinone (UQ). Electron transport generates sufficient energy at three sites (complexes I, III and IV) to pump protons out of the matrix, producing an electrochemical gradient. This gradient is discharged by ATP synthase (complex V) and the energy released used to drive the phosphorylation of ADP to ATP. c, Cytochrome c; e⁻, electron.

DISEASES OF MUSCLE AND THEIR INVESTIGATION

There are a large number of different disorders of muscle, and while our classification includes the main categories (Box 33.1), more comprehensive lists are available (see Karpati et al. in Further reading, below). A detailed description of the clinical features associated with the different types of muscle disease is outside the scope of this chapter, but is discussed in several texts on muscle disease. The clinical features depend upon the age of the

patient and the type of disease. For instance, a child with Duchenne muscular dystrophy will experience difficulty rising from sitting or lying and may have frequent falls. Such problems will prompt the parents to seek advice. In adults, the main forms of presentation are weakness, fatigue and pain. Less commonly, muscle wasting, swelling or twitching of the muscle or a skin rash may be a first symptom. In the genetically determined disorders, the weakness is usually gradually progressive and often follows a characteristic pattern. In other myopathies, there may be associated stigmata, for instance joint disease or skin rash

BOX 33.1 Classification of muscle disease (excluding disorders of motor nerves and the neuromuscular junction) with examples

Non-metabolic, genetically determined myopathies

- Muscular dystrophies
 - Duchenne, Becker
 - Limb girdle types
 - Facioscapulohumeral
 - Others
- Congenital muscular dystrophy/myopathy including central core disease (association with malignant hyperpyrexia)
- Disorders of muscle membrane/myotonic syndromes
 - Myotonic dystrophy type I and II
 - Myotonia congenita
 - Periodic paralyses: hyperkalaemic and hypokalaemic

Trauma to muscle by external agents

- Physical
 - Crush syndrome
 - Ischaemic damage
- Toxic
 - Drugs: steroids, chloroquine, fibrates, HMG-CoA reductase inhibitors (statins), emetine, theophylline (in overdose), zidovudine, snake venoms

Infection

- Viral myositis
- Bacterial myositis

Inflammatory

- Dermatomyositis/polymyositis
- Inclusion body myositis
- Sarcoidosis

Metabolic myopathies

- Muscle disease associated with endocrine disorder
 - Hypo- and hyperthyroidism
 - Hypo- and hyperadrenalism
 - Hyperparathyroidism and osteomalacia
 - Pituitary disorders, e.g. acromegaly
- Genetically determined
 - Disorders of carbohydrate metabolism: myophosphorylase deficiency, acid α -glucosidase deficiency
 - Disorders of fatty acid oxidation: acyl-CoA dehydrogenase deficiency, carnitine palmitoyltransferase deficiency
 - Abnormalities of the respiratory chain: defects of complexes I, III and IV
- Other metabolic myopathies
 - Alcohol myopathies
 - Myopathy with chronic kidney disease
 - Nutritional

Myopathy associated with malignant disease

suggesting a connective tissue disorder; anxiety, sweating and weight loss suggesting hyperthyroidism, or features compatible with high alcohol intake. The muscle pain described by patients with muscle disease may be important in suggesting whether there may be a metabolic cause. For instance, both defects of carbohydrate metabolism and fatty acid oxidation will cause muscle pain associated

with exercise. The pain associated with defects of carbohydrate metabolism occurs during high intensity exercise when glycolysis generates most of the energy required for muscle contraction, whereas defects of fatty acid oxidation cause muscle pain after prolonged exercise at a time when fatty acids are the predominant metabolic fuels.

The clinician must evaluate the clinical features and decide which investigations are appropriate. In many patients with suspected muscle disease, this will involve a combination of biochemical, molecular genetic, neurophysiological and morphological investigations. While many biochemical and genetic studies are performed on blood samples, morphological study and biochemical analyses such as measurement of muscle enzyme activity, require tissue. Muscle biopsy is a relatively simple procedure and there are two main methods: an open biopsy, in which relatively large amounts (0.5–3 g) of muscle can be removed, and a needle biopsy, in which smaller amounts (50–200 mg) are obtained. For most biochemical and histochemical studies, small amounts are sufficient. Morphological changes alone may be sufficient to suggest a diagnosis, for example of Duchenne muscular dystrophy. The diagnosis of metabolic myopathies has been greatly improved by the development of cytochemical techniques that can show, for example, the abnormal storage of glycogen or lipid, or demonstrate the presence or absence of specific enzyme activities in situ. A further development in this area is the use of specific antisera to enable the precise localization (and therefore the presence or absence) of proteins at the cellular level. This technique of immunocytochemistry provides valuable additional information in the investigation of muscle disease.

BIOCHEMICAL INVESTIGATION OF MUSCLE DISEASE

'Routine' biochemical studies

These include the measurement of plasma sodium, potassium, chloride, urea, bicarbonate, glucose, calcium and phosphate, together with simple tests of endocrine function. While not all these tests are necessary for each patient with muscle disease, disturbances of each parameter can result in muscle disease, as shown by the following examples. Severe hypokalaemia associated, for example, with diuretic use or liquorice ingestion, can result in muscle weakness. Renal failure may lead to muscle weakness for several reasons, including electrolyte disturbance and altered calcium metabolism. It must also be remembered that acute muscle necrosis from any cause (e.g. malignant hyperpyrexia, drugs, injury or metabolic myopathy) may itself cause acute kidney injury owing to the tubulotoxic effect of myoglobin. Muscle symptoms are common in endocrine disturbances: hypothyroidism, for example, may be associated with proximal weakness, often with discomfort in the affected muscles.

Plasma creatine kinase activity

The measurement of plasma enzyme activity is important in the diagnosis of muscle disease, and while the activities of several enzymes may be elevated, creatine

kinase (CK) is the most sensitive indicator of muscle damage. Skeletal muscle has the highest CK content of any tissue, more than three times as much as heart or brain, and consequently nearly all CK activity in normal plasma is derived from skeletal muscle. In addition, CK activity is more frequently abnormal than other enzymes in neuromuscular disease and the range of abnormal values is greater.

Human tissues contain three forms of creatine kinase, comprising dimers of the muscle and brain-type subunits, M and B. The combinations are CK-MM, CK-MB and CK-BB. Skeletal muscle contains mostly CK-MM with only a small amount of CK-MB, ranging from 0.2% to 15% of total enzyme activity (mean value of 5–6%). The proportion of CK-MB is higher in type I muscle fibres. Regenerating muscle fibres revert to an embryonic enzyme pattern and have about 40–50% CK-MB. Brain contains only CK-BB while heart muscle contains about 40% CK-MB, the rest being CK-MM. In normal adult plasma, CK activity is almost entirely due to the CK-MM isoform. Muscle damage will increase total CK activity and a proportion will be of the CK-MB type, but this is usually <6%. High plasma CK activity associated with a CK-MB fraction >6% may be associated with myocardial damage; however, there are several other situations in which the plasma CK-MB activity may be high:

- following acute muscle injury or surgical operations in healthy people, especially in those who regularly take vigorous exercise, the plasma CK-MB activity may rise above 6%. This has led to misleading results in marathon runners, acute myoglobinuria and a variety of surgical procedures
- in patients with chronic neuromuscular diseases, there tends to be a higher CK-MB activity owing to the increased percentage of regenerating fibres
- in children younger than early teenage, the percentage of CK-MB in plasma is higher than in adults, in the range of 14–26%. This could complicate the investigation of heart disease in this age group.

The activity of CK in plasma will vary even in healthy subjects and, as noted above, may be significantly elevated following exercise. While CK activity may be raised immediately after exercise, it reaches its peak after 1–2 days. Generally, the more severe the exercise, the higher the total activity and the more delayed the peak. Thus, a 24-fold increase in CK activity has been reported after a 53 mile walk and transient CK values of 10 000–20 000 U/L may be seen after intense training in young men, in whom subsequent investigation fails to detect any abnormality. The effect on CK activity is more pronounced in untrained subjects compared with trained, but generally does not result in greater than a four-fold increase, unless the exercise performed is severe, and returns to normal with resting.

Conditions other than exercise may cause significant increases in plasma creatine kinase activity (Box 33.2). The importance of obtaining a good clinical history and measuring CK activity under basal conditions is, therefore, of paramount importance if unnecessary invasive investigations are to be avoided. Creatine kinase activity in plasma is raised whenever there is necrosis

BOX 33.2 Causes of increased creatine kinase activity other than muscle disease

Muscle origin

- Exercise
- Trauma
- Large muscle mass
- African Caribbean and Indian Asian ethnicity
- Alcohol

Central nervous system disease

- Cerebral infarction or haemorrhage
- Bacterial meningitis
- Head injury

Cardiac disease

- Acute myocardial infarction
- Electrical cardioversion
- Mediastinal radiotherapy

Diseases of other organs

- Pneumonia, lung infarction
- Colonic infarction
- Metastatic carcinoma

Miscellaneous disorders

- Hypothyroidism
- Sepsis
- Shock
- Acute psychosis

Modified from Layzer 1985.

or regeneration of muscle and will, therefore, be elevated in most myopathies. Very high values are found in Duchenne muscular dystrophy and in conditions where severe muscle necrosis occurs, such as acute polymyositis and rhabdomyolysis associated with malignant hyperpyrexia or the metabolic myopathies. In myopathies where muscle destruction is not a major feature, for instance some endocrine myopathies, plasma CK activity may be normal or only mildly elevated. It must be remembered that plasma CK activity may also be elevated in severe neurogenic conditions that include active denervation, and thus it is not diagnostic of muscle conditions alone. Plasma CK activity may be used to follow the progress of myopathic disorders and their response to treatment. For example, the effective treatment of inflammatory muscle disease with steroids and/or other immunosuppressive treatments may be monitored by the fall in CK activity as well as by the improvement in muscle function.

Occasionally, a significantly elevated CK is found unexpectedly. Many of these patients will be found to have muscle disease when formally examined, but there are also individuals with elevated CK who show no signs of neuromuscular disease, so-called 'idiopathic hyperCK-aemia'. Some of these prove to have 'macro CK' – an entity consisting of CK bound to an immunoglobulin (usually IgG) – which is cleared more slowly from the plasma than the unbound enzyme. Endocrine dysfunction, such as hypothyroidism, can elevate CK without obvious muscular dysfunction and it is important to

remember that individuals with malignant hyperthermia can have elevated CK with no other features of neuromuscular disease.

Statin induced elevation of creatine kinase

Drug-induced elevation of CK occurs and, indeed, has become more common since the introduction of the HMG-CoA reductase inhibitors (statins). This class of drug frequently causes myalgia, often with no rise in plasma CK concentration, but only rarely causes muscle disease, although necrotising myopathy and rhabdomyolysis have been reported. Several risk factors that predispose to muscle disease in those taking statins are known, including genetic variants in drug transport molecules and hepatic drug metabolism. Factors that are associated with higher plasma concentrations of statins such as increasing age, female sex, liver or kidney disease, untreated hypothyroidism and drugs that inhibit the cytochrome P450 system, increase the risk of myalgia and rhabdomyolysis. Interfering drugs that are of particular concern include cyclosporin, fibrates, calcium channel blockers, protease inhibitors and warfarin; grapefruit juice also has a similar effect. Pravastatin is not metabolised by the cytochrome P450 system so may be a safer choice than other statins if combination treatment with these drugs is required. Low plasma concentrations of 25-hydroxycholecalciferol are associated with myalgia and may exacerbate statin associated muscle pain. There is some evidence that patients given vitamin D replacement before starting a statin are at lower risk of developing symptoms.

There is a strong association between statin induced myalgia and a common variant of the gene *SLCO1B1*, which codes for a transporter that mediates the hepatic uptake of statins apart from fluvastatin. In one study, >60% of cases could be attributable to the variant, which impairs statin uptake. Cyclosporin is thought to inhibit the transporter and can therefore increase plasma concentrations of most statins, even of those that are not metabolised by the cytochrome P450 system.

The underlying mechanism of statin toxicity is still debated, but theories include alterations in myocyte membrane cholesterol content, depletion of isoprenoids that control myofibre apoptosis and depletion of coenzyme Q₁₀ (ubiquinone). Hydrophilic statins such as pravastatin and rosuvastatin may be less myotoxic than other more lipophilic statins such as atorvastatin, simvastatin and lovastatin, possibly because of decreased penetration into muscle cells.

In most patients with statin-induced myalgia, the CK does not rise dramatically, if at all, and a minor rise may simply require monitoring and not cessation of treatment. Strategies for improving the tolerability of statin treatment include changing to a hydrophilic statin, reducing the frequency of doses to once or twice weekly, or supplementation with vitamin D and/or coenzyme Q₁₀. However, much of the evidence of benefit of dosage variations and supplements has been anecdotal and there are few data on the cardiovascular effects of such strategies.

Persisting elevation of CK following withdrawal of statin treatment should be investigated, as there is more likely to be an underlying muscle disorder. The use of

statins in patients with known muscle disease should be avoided if possible, although in individuals where there is a strong case for treatment, monitoring CK activity to detect progressive change is an alternative.

Other enzymes measurable in plasma

The activities of aldolase, lactate dehydrogenase and pyruvate kinase are also raised in destructive myopathies, but measurement of these enzymes provides no additional information to that given by CK measurement. Similarly, carbonic anhydrase III and a muscle-specific enolase, two enzymes that are specific to skeletal muscle, are also elevated in muscle disease, but the patterns of abnormality are similar to that of CK. Alanine aminotransferase (ALT) and aspartate aminotransferase (AST) are also raised in the presence of muscle destruction and rise in parallel with CK. Measurement of plasma γ -glutamyl transferase activity may help to determine the tissue origin of high aminotransferases, as it is normal in muscle disease but usually raised with liver disease.

Myoglobinuria

In the presence of muscle damage, the muscle pigment myoglobin leaks into blood and urine. While myoglobin is a sensitive indicator of muscle destruction, it is neither significantly more sensitive nor more specific than CK. Myoglobin gives a positive reaction on urine dipstick testing for blood, which can be misinterpreted as being due to haematuria. In conditions in which there is severe muscle necrosis, rhabdomyolysis may occur. Causes include mechanical trauma, myotoxic agents (animal poisons, numerous drugs and alcohol), extremes of ambient temperature, infections and inherited muscle diseases such as malignant hyperpyrexia, some types of muscular dystrophy and defects of mitochondrial energy, carbohydrate and lipid metabolism. Myoglobin may be deposited in nephrons and cause acute kidney injury. The risk of injury appears to be associated with acute elevations of myoglobin.

INVESTIGATION OF MUSCLE DISEASE

Non-metabolic, genetically determined myopathies

The investigation of muscular dystrophies and other non-metabolic myopathies is largely dependent on morphological and/or genetic analyses. For example, the X-linked allelic disorders Duchenne and Becker muscular dystrophies may be diagnosed by identifying the defect at the genetic level; a deletion of the dystrophin gene is found in around 65% of Duchenne boys. The same is true for facioscapulohumeral muscular dystrophy, where a deletion of telomeric DNA on the long arm of chromosome 4 is seen in >95% of cases. In the limb girdle group of muscular dystrophies, the same phenotype is produced by defects in very different proteins (and their corresponding genes) such that muscle biopsy and immunocytochemical studies are usually required.

In membrane disorders, such as the myotonic dystrophies and congenital myotonias, genetic analysis is paramount. In contrast, the periodic paralyses, which are also caused by ion channel dysfunction, are associated with electrolyte changes that are important to define. In hypokalaemic periodic paralysis, the patient may experience episodes of weakness that last from hours to days and which can be provoked by carbohydrate ingestion. Serum potassium measured at the onset of weakness is low, but may normalize quickly despite persisting symptoms. In hyperkalaemic periodic paralysis, the potassium concentration is elevated and there is also a form in which potassium appears not to change. In certain populations, particularly in Asia, the combination of hypokalaemic paralysis and hyperthyroidism is found. Treatment of the underlying thyroid dysfunction cures the potassium disturbance and the muscle disease.

Malignant hyperpyrexia is a rare syndrome, in which there is a rapid rise of body temperature (up to 1 °C every 5 min), lactic acidosis, muscle rigidity, hyperkalaemia, very high plasma CK activity and myoglobinuria. The reaction may be triggered by several inhalational anaesthetics (e.g. halothane) and succinylcholine. The disorder has been linked to various different genetic defects including mutations in the ryanodine receptor gene. (Defects in this gene also give rise to the muscle disease, central core disease.) Plasma CK activity can also be increased in patients at risk. Diagnosis is dependent on *in vitro* testing of muscle strips for an abnormal contracture response to halothane, caffeine or a combination of both agents.

Metabolic, genetically determined myopathies

Whereas the biochemical investigations so far described are either general indicators of muscle damage or directed at systemic conditions in which muscle involvement occurs, the following section deals with specific investigations of genetically determined metabolic muscle disease. It is in this area that biochemical studies are of paramount importance. The three main areas to be described are disorders of carbohydrate metabolism, defects of the respiratory chain and defects of fatty acid oxidation. The discussion of each will be divided into sections dealing with the investigations performed in blood and/or urine, muscle tissue analyses (e.g. histochemical studies, when appropriate) and specific biochemical investigations.

Disorders of carbohydrate metabolism

Disorders of carbohydrate metabolism can be divided into two main groups:

- failure to metabolize glucose, either owing to impaired mobilization or breakdown of glycogen (e.g. myophosphorylase deficiency) or impaired glycolysis (e.g. phosphofructokinase deficiency)
- disorders resulting from lysosomal storage of glycogen (α -acid maltase deficiency) or from the deposition of abnormal polysaccharide (branching enzyme deficiency). For a detailed discussion of the

clinical and biochemical features, the reader is referred to the review by Chen (2001) (see Further reading, below).

Dynamic/functional tests. In diseases related to failure to metabolize glucose, the symptoms are related to the lack of energy (ATP) under conditions where muscle relies on glycolysis to meet this demand. The main clinical manifestations are cramp-like muscle pain and, occasionally, the appearance of contractures. The inability to derive energy from glycolysis is easily demonstrated by making the patient perform muscular work and measuring the rise of blood lactate. This is usually done under ischaemic conditions, but is equally effective without ischaemia (see [Appendix 33.1](#)).

Histochemistry. Histochemical methods are used to study both the amount of glycogen and several of the enzymes of glycogen mobilization and glycolysis. Glycogen is usually demonstrated by some variant of the periodic acid Schiff (PAS) technique. Myophosphorylase, phosphofructokinase, myoadenylate deaminase and lactate dehydrogenase activities are all detectable cytochemically, and absence of enzyme activity is diagnostic in clinical terms. The finding of glycogen storage with grossly elevated membrane-bound acid phosphatase activity is virtually diagnostic of acid α -glucosidase deficiency.

Biochemical investigations. The enzymes of glycogenolysis and glycolysis are all highly active, and homogenates prepared from needle-biopsy specimens are usually adequate for analysis. The measurement of these enzymes has been reviewed elsewhere (see Further reading, below) and only a brief summary is given here.

Acid α -glucosidase (acid maltase) is measured by monitoring glucose release from maltose or glycogen at acid pH or by using 4-methylumbelliferyl α -glucoside as substrate and measuring the fluorescence of 4-methylumbelliferone released. The process has been simplified by the availability of a sensitive immune capture technique and activity can now be measured in dried blood spots saved on filter paper. The simplest technique for the assay of **amylo-1, 6-glucosidase** activity is based on the fact that the hydrolytic activity of this enzyme is partly reversible; activity may be followed by measuring the incorporation of [14 C] glucose into glycogen in the presence of a muscle homogenate. **Branching enzyme** α -1: α -1,4-glucan:1,4-glucan-6-glucosyltransferase activity depends upon the stimulation of the rate of glycogen synthesis from glucose 1-phosphate catalysed by phosphorylase a (non-physiologically): activity is followed by the release of inorganic phosphate. The reaction catalysed by **phosphorylase a** is reversible, although, *in vivo*, the concentration of glucose 1-phosphate is at least two orders of magnitude too low to allow glycogen synthesis. The enzyme can therefore be measured in either direction. **Phosphorylase kinase** is measured by the rate of ATP-dependent conversion of phosphorylase a to phosphorylase b. **Phosphofructokinase** activity is measured in muscle homogenates by coupled enzyme assay in which the fructose 1,6-bisphosphate formed is converted to dihydroxyacetone phosphate, and this is converted to glycerol 3-phosphate with the oxidation of

NADH. **Phosphoglycerate kinase** and **phosphoglycerate mutase** are measured in muscle homogenates by coupled enzyme assays. **Lactate dehydrogenase** is measured by a standard spectrophotometric assay.

Defects of the respiratory chain

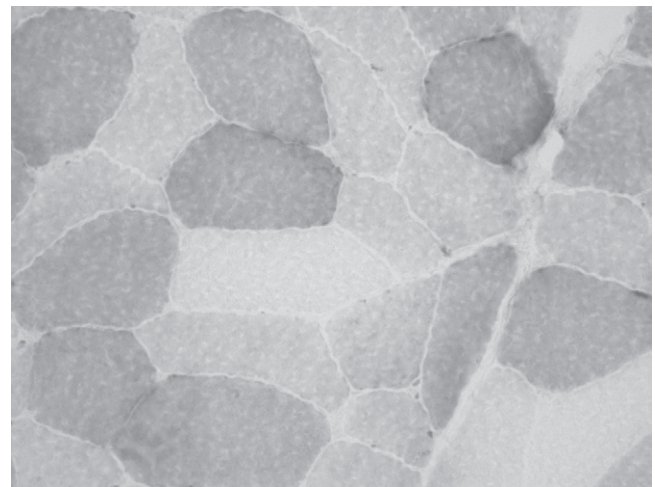
These disorders were first identified in muscle, but are now known to produce a wide range of clinical disorders often involving the central nervous system (see Further reading, below). The respiratory chain is composed of five complexes (Fig. 33.7) and defects have been identified in each of them as well as in many of the proteins required to assemble and maintain this pathway and the mitochondrial genome that encodes some of the components.

Dynamic/functional tests in blood. Elevated plasma lactate, often associated with high concentrations of pyruvate and alanine, is a common, but not universal feature of respiratory chain dysfunction. In some patients, resting lactate is normal, but increases to very high concentrations during aerobic exercise. Aerobic exercise tests can, however, be difficult to perform in children and patients with significant physical disability. Examination of the CSF lactate is of value in patients who have predominantly CNS symptoms since the lactate concentration may be inappropriately high in this compartment, but normal or near normal in blood. In patients with the Kearns–Sayre syndrome, elevated total protein concentrations are found in CSF, and this is one of the diagnostic criteria of the disorder. Recently, a fibroblast growth factor, FGF21, was found to be elevated in plasma in mitochondrial disease and appears to be a better biomarker for mitochondrial muscle disease than metabolites such as lactate.

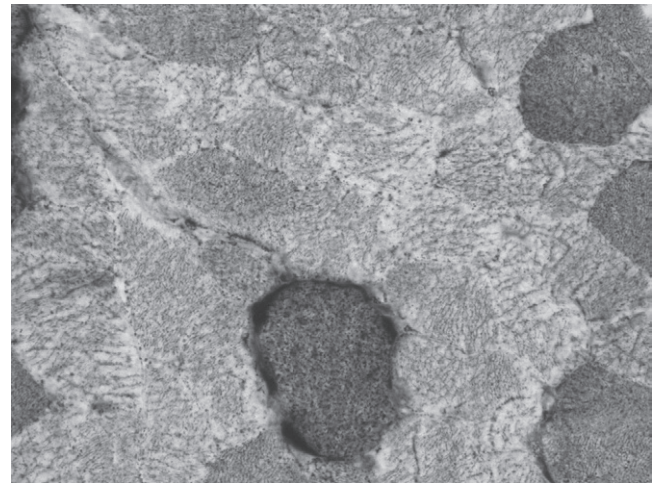
Histochemistry. The activity of several mitochondrial enzymes can be demonstrated cytochemically. Cytochrome *c* oxidase (COX) and succinate dehydrogenase (SDH) are the most specific for the respiratory chain, but NADH oxidation and ATPase activity may also give non-specific clues to the presence of a disorder. The morphological hallmark of respiratory chain disease is the ragged red fibre originally described using the Gomori trichrome stain. The finding of COX positive and deficient muscle fibres (a mosaic) is highly suggestive of a defect in mitochondrial DNA (mtDNA), and this can easily be shown using COX and SDH stains in combination (Fig. 33.8). In this case, histochemical analysis is sufficient and the next step will be genetic analysis of mtDNA. Detailed biochemical analysis is still necessary, however, particularly for the investigation of complexes I and III. Newer techniques that combine gel electrophoresis, ‘in gel’ activity measurement and immunoblotting are also available and will certainly play a greater role in the diagnostic process in the future.

Biochemical investigations

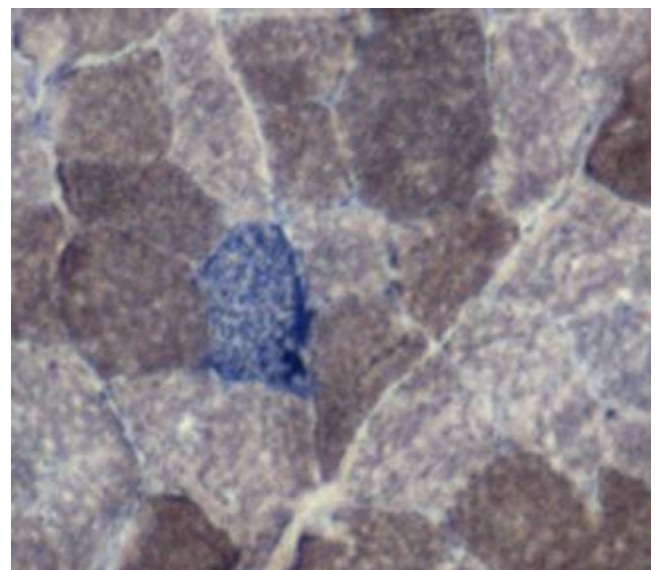
Measurement of mitochondrial oxidations. The complete oxidation of metabolic fuels with concomitant synthesis of ATP can be studied in various ways, including overall flux from substrate to water, flux through various segments and the activities of individual components



A



B



C

FIGURE 33.8 ■ Histochemical demonstration of: (A) cytochrome *c* oxidase (complex IV) and (B) succinate dehydrogenase (complex II) activity. The reactions specifically identify respiratory chain complexes and the combination (C) permits the easy detection of muscle fibres lacking cytochrome *c* oxidase activity (light grey fibres). (Photographs courtesy of Dr B Krossnes.)

of the respiratory chain. These will be discussed below. In many instances, it is necessary to isolate mitochondria from tissue (usually muscle). This is a relatively straightforward, but critical, step, since the outcome of all further work will depend on the quality of the mitochondrial fraction prepared.

Flux through the whole respiratory chain can be studied either by measuring oxygen consumption or the production of ATP. Polarography measures the rate of change in oxygen concentration in solution, and since oxygen is the final electron acceptor of the respiratory chain, this provides direct measurement of activity. The working volume of the electrode determines how much mitochondrial protein is required and since human muscle biopsies are small, the smaller the volume the better. Several parameters may be assessed using polarography. First, the quality or state of the mitochondria may be inferred from the respiratory control ratio (RCR). This is the ratio of oxygen consumption in the presence of ADP to that when all ADP has been converted to ATP. The RCR in fact reflects the structural integrity of the mitochondria, since any damage to the inner membrane will dissipate the electrochemical gradient and thus the ability to synthesize ATP. The RCR in human skeletal muscle mitochondrial fractions is usually between 2 and 4. Flux through the respiratory chain is often expressed in terms of oxygen consumed per milligram of protein in the presence of ADP. It is also possible to measure ATP formed during respiration, either directly or by linking to a luminescent marker.

Flux through segments of the respiratory chain can be measured spectrophotometrically using artificial electron acceptors whose light absorption changes on reduction/oxidation. A commonly used acceptor is potassium ferricyanide. NADH produced by the oxidation of pyruvate, oxoglutarate or glutamate, yields electrons that enter the respiratory chain via complex I. Ferricyanide accepts electrons from cytochrome *c* with the result that it is possible to assess flux through the segment containing complex I, ubiquinone and complex III (see Fig. 33.7). Alternatively, using succinate as a substrate, electrons enter via succinate dehydrogenase that is part of the respiratory chain complex II. The ferricyanide-linked assay will then measure flux from complex II to cytochrome *c*.

Measurement of activity of individual respiratory chain complexes. The activity of these complexes is best measured in mitochondrial fractions. Where small amounts of tissue are available, partial purification is used.

Complex I activity is measured by determining rotenone-sensitive NADH-ubiquinone oxidoreductase activity. Low molecular weight, water-soluble ubiquinone or a ubiquinone analogue can be used as an electron acceptor. This assay requires that the mitochondria are made permeable to enable the NADH to reach its binding site on the inner aspect of the inner mitochondrial membrane. NADH oxidation is measured by the decrease in absorbance at 340 nm before and after the addition of rotenone. The use of rotenone, a specific complex I inhibitor, is essential since there are other enzymes capable of oxidizing NADH.

Complex II activity is measured by following the reduction of ubiquinone by succinate either directly or by

linking the reaction to an artificial electron acceptor such as dichlorophenol-indophenol. The succinate dehydrogenase component of complex II can be further studied by using phenazine ethosulphate rather than ubiquinone as the intermediate electron acceptor.

Complex III activity is measured by following the reduction of cytochrome *c* by ubiquinol. Ubiquinol is synthesized from ubiquinone by reduction with dithionite followed by organic extraction. Further studies of part of complex III can be made by measuring the transhydrogenase reaction, which gives information on the *b* cytochromes.

Complex IV activity is measured by following the oxidation of reduced cytochrome *c*. Reduced cytochrome *c* is prepared by addition of ascorbate, which is removed by gel filtration. Polarographic measurement of complex IV activity is also widely used. In this reaction, ascorbate and tetramethyl-*p*-phenylenediamine dihydrochloride (TMPD) are used as artificial electron donors to cytochrome *c*.

Complex V is rarely measured in patients with suspected mitochondrial disease, but several methods to measure oligomycin-sensitive ATPase activity have been used for mitochondria from other species.

Molecular biology techniques. Mitochondria are unique in that they contain their own DNA (mtDNA). This small genome encodes 13 polypeptides (seven that go into complex I, one in complex III, three into complex IV and two into complex V) and 24 RNA species that participate in intramitochondrial protein synthesis. Disorders caused by mtDNA mutations are highly variable, but there are some classic syndromes that are easily recognizable. For example, Kearns–Sayre syndrome and chronic progressive external ophthalmoplegia that can be caused by gene rearrangements (most often deletions); Leber hereditary optic neuropathy is due to a point mutation in complex I protein-coding genes, and myoclonus, epilepsy with ragged red fibres (MERRF) and myopathy, encephalopathy, lactic acidosis and stroke (MELAS) are due to point mutations in different mitochondrial genes. Since the mitochondrial genome is inherited exclusively from the mother, disorders in which maternal inheritance occurs gave the first clue to the possibility of mtDNA mutations.

Since the first mtDNA mutations were described in 1988, more than 150 different mutations have been found including both rearrangements (deletions or duplications) and point mutations. In the last few years, mutations in nuclear encoded genes affecting mitochondria have also been identified and this group is expanding. As stated earlier, the finding of a mosaic of COX positive and deficient fibres is highly suggestive of a defect involving mtDNA. This can, however, either be a primary mtDNA mutation, as, for example, in the syndrome of MELAS, or a secondary mtDNA defect caused by a nuclear genetic defect in a protein that has a role in mtDNA maintenance.

Defects of fatty acid oxidation

Muscle symptoms associated with a defect of fatty acid oxidation can occur alone or as part of a systemic abnormality. Symptoms may arise acutely or more chronically

and affected individuals are at risk of developing cardiomyopathy, hepatic dysfunction, sudden infant death and hypoglycaemic coma.

Dynamic/functional tests

Intermediary metabolites and metabolic fuels in blood. If a systemic defect is present, the inter-relationship between metabolic fuels is abnormal. In healthy subjects during stress or fasting, lipolysis is stimulated and free fatty acids are released into the circulation. Following uptake by tissues, fatty acids are converted to their acyl-CoA esters and oxidized by β -oxidation to generate acetyl-CoA. Acetyl-CoA may be further oxidized by the citrate cycle or converted to ketone bodies in the liver. Circulating ketone bodies are an important source of energy during stress or fasting and are readily oxidized by extrahepatic tissues, in particular muscle and brain. When fatty acid oxidation is impaired, the blood concentrations of free fatty acids increase, but there is no concomitant increase in the concentration of ketone bodies. Energy must, therefore, be derived from carbohydrate metabolism as long as glycogen stores last. Once these are depleted, hypoglycaemia can develop and may lead to permanent tissue damage or death. It is clear, therefore, that fasting is potentially dangerous for patients with fatty acid oxidation defects and should be avoided.

Measurement of plasma, tissue and urine carnitine concentrations. Patients with defects of fatty acid oxidation frequently have abnormalities of carnitine metabolism. Partial β -oxidation will continue to generate chain-shortened acyl-CoA ester intermediates that combine with carnitine in a reaction catalysed by carnitine acyltransferases. These carnitine esters are transported out of the mitochondrial matrix, resulting in a higher percentage of carnitine in the acylated form in blood and urine from patients compared with healthy subjects. In addition, in the presence of a defect in β -oxidation, there will be abnormal excretion of specific acylcarnitines, e.g. octanoylcarnitine in medium chain acyl-CoA dehydrogenase deficiency (MCAD). Secondary carnitine deficiency can occur in these individuals, although the mechanism remains uncertain, as total urinary excretion of acylcarnitines is not increased. Carnitine concentrations, free and acylated, are usually measured using tandem mass spectrometry. Since carnitine metabolism is perturbed by most defects of fatty acid oxidation, these measurements rarely help in making a specific diagnosis.

Mass spectroscopic measurement of specific acylcarnitines is a screening technique that is now widely used. This can be performed on dried blood spots (e.g. on a Guthrie card) and allows the detection of specific intermediates that are identified using tandem mass spectrometry. The metabolism of fatty acids proceeds by a chain shortening series of reactions that produce acetyl-CoA. In the presence of a block in this process, chain-shortened fatty acids will accumulate and these are esterified with carnitine, forming acylcarnitines that can be detected in blood. [Figure 33.9](#) shows typical profiles for three different defects that can be identified using this technique in dried blood samples. Owing to the frequency of MCAD

in certain populations and its association with sudden infant death, this technique is increasingly being used for screening of neonates.

Measurement of dicarboxylic acids and acylglycines in urine. When mitochondrial β -oxidation is impaired, dicarboxylic acids, e.g. adipic, suberic and sebacic, are generated by the partial oxidation of fatty acids in the endoplasmic reticulum (ω -oxidation) and by β -oxidation in peroxisomes. These dicarboxylic acids are excreted in the urine, and the pattern of the organic aciduria, detected by gas chromatography, may be helpful in the diagnosis of a defect of fatty acid oxidation. Acylglycines are formed in the liver by the combination of acyl-CoA esters and glycine, a reaction catalysed by glycine-N-acylase. Glycine conjugates are excreted in the urine and can be detected by gas chromatography. An isotope dilution analysis of urinary acylglycines has been introduced, which seems to be a very sensitive test for defects of medium chain acyl-CoA dehydrogenase deficiency.

Specific biochemical investigation

Measurement of flux through β -oxidation. Measurement of flux is an important investigation in defects of fatty acid oxidation. There are, however, many problems in accurately determining the flux in muscle preparations for diagnostic purposes. One of the most frequently used methods measures $^{14}\text{CO}_2$ formed by the oxidation of ^{14}C -labelled fatty acids. This method is inaccurate since $^{14}\text{CO}_2$ is only released by the further oxidation of acetyl-CoA by the citrate cycle, which is incomplete and not necessarily proportional to the flux through β -oxidation. Measurement of ^{14}C -labelled acid-soluble products is a much better indicator of flux, but this method is not suitable for ^{14}C -labelled medium and short chain fatty acids. An alternative approach is to determine $^3\text{H}_2\text{O}$ formed during the oxidation of ^3H -labelled fatty acids. Unfortunately, this technique is also limited because of the relatively few ^3H -labelled fatty acids available. Acylcarnitine profiles can be studied in fibroblasts using the same technique as for dried blood spots.

Measurement of carnitine transport and enzyme activity. Primary carnitine deficiency, due to a defect of carnitine transport, can be investigated by measuring the uptake of labelled carnitine into fibroblasts. Carnitine palmitoyltransferase activity is measured radiochemically. A number of different spectrophotometric and fluorometric techniques have been used to measure acyl-CoA dehydrogenase activities and the methods of choice are a dye reduction assay and the fluorometric ETF-reduction assay. Electron transfer flavoprotein activity can be measured by a dye reduction assay or by a catalytic assay using a physiological electron donor, reduced medium chain acyl-CoA dehydrogenase and the physiological electron acceptor ETF dehydrogenase. Electron transfer flavoprotein dehydrogenase activity can be measured by two different methods. One measures the NADH-ETF reductase activity anaerobically following the reduction of the ETF_{ox} flavin to the semiquinone. The other measures the co-proportionation of oxidized- and two-electron reduced to one-electron reduced ETF

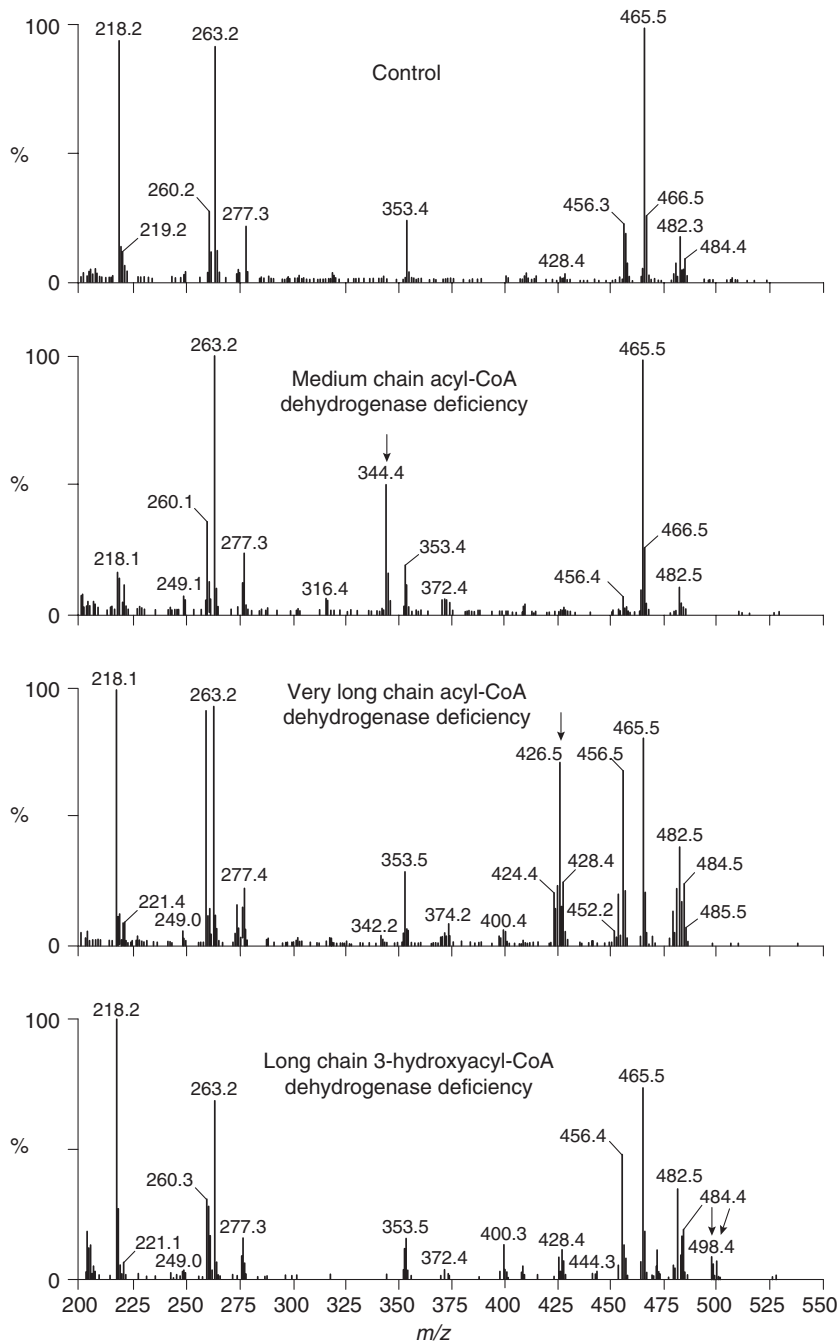


FIGURE 33.9 ■ Analysis of acylcarnitine intermediates using mass spectrometry. Acylcarnitine profile in dried blood spot from a healthy subject (upper panel) and patients with defects of fatty acid oxidation. The numbers above the peaks are m/z values that correspond to the molecular ions of butylated acylcarnitine species: free carnitine (m/z 218), free carnitine internal standard (m/z 221), C2-carnitine (m/z 260), C2-carnitine internal standard (m/z 263), C3-carnitine (m/z 274), C3-carnitine internal standard (m/z 277), C8-carnitine (m/z 344), C8-carnitine internal standard (m/z 353), C12-carnitine (m/z 400), C14:1-carnitine (m/z 426), C14-carnitine (m/z 428), C16-carnitine (m/z 456), C16-carnitine internal standard (m/z 465), C18:1-carnitine (m/z 482), C18-carnitine (m/z 484), 3-OH-C18:1-carnitine (m/z 498), 3-OH-C18-carnitine (m/z 500). Major diagnostic analytes are shown by arrows. For example, in medium chain acyl-CoA deficiency, medium chain (C8)-carnitine esters accumulate (m/z 344). In very long chain acyl-CoA dehydrogenase deficiency, C14 and longer chain fatty acids accumulate, while in long chain 3-hydroxyacyl-CoA dehydrogenase deficiency (bottom panel), long chain hydroxy forms of the carnitine esters accumulate.

semiquinone by following the disappearance of the ETF_{ox} flavin fluorescence. Enoyl-CoA hydratase activity is measured by following the hydration of the double bond of 2-enoyl-CoA esters. 3-Hydroxyacyl-CoA dehydrogenase activity is measured in the reverse direction following the oxidation of NADH or in the forward

direction by linking the reaction with 3-oxoacyl-CoA thiolase. 3-Oxoacyl-CoA thiolase activity is measured by following the decrease in absorbance resulting from the disappearance of a Mg^{2+} -enolate complex using a long chain 3-oxoacyl-CoA ester and acetoacetyl-CoA as substrates.

CONCLUSION

The investigation of muscle disease requires a combination of clinical investigation, morphological, biochemical and molecular genetic studies. Clinicians must recognize the characteristic patterns of weakness and important associated clinical features and then decide which investigations are appropriate. Biochemical studies will play an important part in the investigation, particularly in those patients with suspected metabolic myopathies. Many of the routine biochemical studies will be available at all hospitals, but the specialized tests to investigate metabolic myopathies are best performed in the few centres with a research interest in muscle disease.

ACKNOWLEDGEMENTS

I am grateful to my colleagues Mike Cullen, Margaret Johnson, Mori Pourfarzam and Bård Krossnes for giving me their illustrations and their experience and to Petter Sanaker for reading through the manuscript.

Further reading

- Berardo A, DiMauro S, Hirano M. A diagnostic algorithm for metabolic myopathies. *Curr Neurol Neurosci Rep* 2010;10:118–26.
This review describes the clinical features encountered in metabolic myopathies, particularly mitochondrial disease.
- Chen Y-T. Glycogen storage diseases. In: Scriver CR, Beaudet AL, Sly WS et al. editors. 8th ed. New York: McGraw-Hill; 2001 Chapter 71.
This describes the clinical features and the biochemical investigations of defects of carbohydrate metabolism.
- Edwards RHT, Young A, Wiles CM. Needle biopsy of skeletal muscle in the diagnosis and clinical study of muscle function and repair. *N Engl J Med* 1980;302:267–71.
- Jackson MJ, Schaefer J, Johnson MA et al. Presentation and clinical investigation of mitochondrial (respiratory chain) disease: a study of 51 patients. *Brain* 1995;118:339–57.
- Johnson MA. Skeletal muscle. In: Filipe MI, Lake BD, editors. *Histochemistry in pathology*. Edinburgh: Churchill Livingstone; 1990. p. 129–57.
The morphological and histochemical evaluation of muscle disease is described in this chapter.
- Karpati G, Hilton-Jones D, Bushby K, Griggs RC, editors. *Disorders of voluntary muscle*. 8th ed. Cambridge: Cambridge University Press; 2009.
This is a comprehensive review of muscle disease and is an excellent reference volume.
- Layzer RB. *Neuromuscular manifestations of systemic disease*. Philadelphia: F A Davis; 1985.
This book describes the common involvement of muscle in systemic disease.
- McFarland R, Taylor RW, Turnbull DM. A neurological perspective on mitochondrial disease. *Lancet Neurol* 2010;9:829–40.
A review of the neurological manifestations of mitochondrial (respiratory chain) disease.
- Nanji AA. Serum creatine kinase isoenzymes: a review. *Muscle Nerve* 1983;6:83–90.

This review describes the different forms of creatine kinase and the relevance of this investigation in various muscle diseases.

Sherratt HSA, Watmough NJ, Johnson MA, Turnbull DM. Methods used for the study of normal and abnormal skeletal muscle mitochondria. *Methods Biochem Anal* 1988;33:243–335.

The review describes the investigation of skeletal muscle mitochondrial disease.

Warren JD, Blumbergs PC, Thompson PD. Rhabdomyolysis: a review. *Muscle Nerve* 2002;25:332–47.

A comprehensive review.

APPENDIX 33.1: THE FOREARM EXERCISE TEST

This consists of measuring lactate production following 1 min of exercise, which can be performed under ischaemic conditions or without. The use of ischaemia in patients with glycogen breakdown defects such as McArdle disease can be associated with rhabdomyolysis and should be avoided. The test is otherwise the same. If used, ischaemia is induced in the forearm by the use of a sphygmomanometer cuff placed around the upper arm and inflated to just above systolic pressure (*note*: a reliable cuff is essential since gradual loss of pressure will allow the entry of fresh blood).

The subject is asked to squeeze a (spare) sphygmomanometer bulb once every second for 1 min and then rest with the cuff still inflated for 1 min to allow lactate to leak out of the forearm muscles. The cuff is then released. Blood samples are taken prior to inflating the cuff and immediately following release, and at intervals (e.g. 2, 5, 7 and 12 min) after exercise. The pre-exercise blood sample must be withdrawn without stasis so a small, indwelling cannula (e.g. 21-G butterfly) is placed in an antecubital vein. Should stasis be required to place this, 10 min should be left before taking the pre-exercise sample. Samples taken into anticoagulants (e.g. fluoride oxalate) must be analysed within 30 min. Alternatively, samples can be mixed with ice-cold perchloric acid.

The normal response is for lactate to rise in the first sample after ischaemic exercise to between three and five times pre-exercise concentrations and then gradually decline. Healthy subjects perform this test easily, but patients with disorders of glycolysis often have to stop the exercise owing to pain and/or cramp. In disorders of glycogen mobilization or glycolysis, little or no rise in venous lactate will be seen. Since myoadenylate deaminase deficiency gives a similar clinical picture to myophosphorylase deficiency, ammonia concentration should be measured simultaneously with lactate. In myoadenylate deficiency, lactate rises normally, but there is no (normal) rise in ammonia. This provides, moreover, an internal control for the investigation in most instances, since failure of lactate to rise can also be due to poor effort during ischaemic exercise.

Investigation of cerebrospinal fluid

Geoffrey Keir • Carrie Chadwick

CHAPTER OUTLINE

INTRODUCTION 660

CEREBROSPINAL FLUID PHYSIOLOGY 661

Formation 661

Composition 661

Analysis of cisternal or ventricular fluid 661

INVESTIGATIONS RELEVANT TO PHYSIOLOGY AND PATHOPHYSIOLOGY 661

Sampling and pressure 661

Appearance 661

Cells 661

Glucose 662

Lactate 662

Proteins 662

Brain-specific proteins 665

Cerebrospinal fluid oto- and rhinorrhoea 666

Haem pigments and ferritin 666

Enzymes in CSF 667

Markers of inflammation 667

Non-biochemical investigations 668

BIOCHEMICAL INVESTIGATIONS IN CNS DISORDERS 669

Acute infections 669

Chronic infections 669

Haemorrhage and obstruction 670

Inherited metabolic diseases 670

Malignancy 670

Dementia 670

Cerebrospinal fluid analysis in demyelinating diseases 671

CONCLUSION 671

INTRODUCTION

Spinal paracentesis (lumbar puncture) was introduced at the end of the 19th century for the therapeutic relief of raised intracranial pressure secondary to tuberculous meningitis. Within a few years, however, it was being used to sample lumbar cerebrospinal fluid (CSF) to aid clinical diagnosis. Analyses that were undertaken in these early days included microscopy for cells and chemical tests to measure reducing substances, protein and chloride. Even today, counting cells and measuring total protein and glucose remain the mainstay analyses routinely carried out on CSF specimens. Yet, modern analytical techniques make possible the quantitative and qualitative detection of a vast range of biochemical substances. It is the purpose of this chapter to assess critically the role of laboratory studies of CSF in the investigation of central nervous system (CNS) disorders. The results of these analyses must be considered both in the clinical context and in conjunction with the findings of other investigations, including neuroimaging and neurophysiological measurements.

A few points should be kept in mind when handling CSF. Despite the fact that well-documented procedures exist that will minimize discomfort, including that of post-lumbar puncture headache, the anticipation of a lumbar puncture (LP) is something that causes distress

to many patients. This is not helped by the fact that, occasionally, even when carried out by an experienced practitioner, performing a satisfactory LP can prove difficult and uncomfortable and is not without the risk of epidural or subdural bleeding or infection. In many instances, only one spinal tap can be carried out. One reason is that the introduction of a needle into the lumbar sac to sample the CSF can lead to blood contamination of the remaining CSF. This can invalidate the taking of further samples for several days, the exact time being dependent upon the analyte being studied and the degree of blood contamination. There are also ethical reasons against repeating an invasive procedure for trivial reasons. Furthermore, in the very young, the volumes of CSF that can safely be taken are small. Collectively, this places a heavy burden of responsibility on laboratories to make sure that the maximum relevant information is achieved from the available fluid, and that none is wasted or prematurely discarded. Also, the analysis of CSF is multidisciplinary, and it behoves laboratories to ensure that there is the necessary cooperation among pathology specialties. It is strongly recommended that laboratories produce guidelines for doctors indicating sample requirements, and establish procedures to ensure that all CSF samples surplus to immediate analytical requirements are correctly stored. Like all body fluids, CSF is a potentially hazardous material and must be handled appropriately.

CEREBROSPINAL FLUID PHYSIOLOGY

Formation

In the normal adult, the total volume of CSF is about 150 mL. Cerebrospinal fluid is formed at a rate of approximately 500 mL/day, so the fluid is typically exchanged about four times daily (every 6 h). Infants and children have smaller CSF volumes that range from 30–60 mL in a neonate to 100 mL in a pre-teen.

While the choroid plexuses that line the ventricles are the major site of CSF production, as much as 30% of CSF is formed by fluid shifts across various vascular beds within the CNS, such as the cerebral capillaries and dorsal root ganglia. Unlike choroidal CSF, interstitial fluid forms throughout the brain and spine, and then diffuses into the CSF.

Cerebrospinal fluid formed in the lateral ventricles passes through the third and fourth ventricles into the cisterna magna, from where it circulates out into the cerebral and spinal subarachnoid spaces. The subarachnoid space lies between the two leptomeninges, the outer (arachnoid) mater and inner (pia) mater, which cover the whole brain and spinal canal down to the level of the second sacral vertebra. The flow of CSF from the cisterna magna is mainly upward and outward over the cerebral hemispheres to the main site of reabsorption through the arachnoid villi, which drain into the major dural sinuses. Flow down the cul-de-sac of the spinal cord is relatively sluggish. Thus, while the average turnover time of CSF is ~6 h, radiolabelled plasma albumin can continue to equilibrate with lumbar CSF for 1–2 days.

Composition

Solutes enter the CSF by a variety of processes, including active transport via specific transporter mechanisms present in choroidal epithelial cells, facilitated diffusion and passive diffusion.

Cerebrospinal fluid constituents may be derived from plasma, from the metabolic activities of cells normally present in the CNS or from cells and organisms present in the CSF as a result of pathological processes.

Analysis of cisternal or ventricular fluid

While the most common fluid obtained for analysis is lumbar CSF, there are occasions when CSF is taken from other regions, for example from one of the lateral ventricles (usually via a ventriculo-peritoneal or ventriculo-extracorporeal shunt), the 4th ventricle (by cisternal puncture) and even from around the cortex. It is extremely important that the anatomical site of the fluid is recorded, as the reference ranges for many constituents vary according to the anatomical source of the CSF.

INVESTIGATIONS RELEVANT TO PHYSIOLOGY AND PATHOPHYSIOLOGY

Sampling and pressure

Details of CSF sampling techniques are well described in standard textbooks of practical medical procedures. The

initial hydrostatic pressure is between 80 and 180 mmH₂O when the patient is in the lateral recumbent position. Ideally, four sequential fractions of CSF should be collected into sterile polypropylene containers. In an adult, a total of 10–12 mL of fluid should be withdrawn, but as little as 3 mL or less in the case of a neonate. The containers must be numbered in the order of collection. Typically, fraction 1, taken into a fluoride tube, is used for glucose (this fraction may also be suitable for protein measurement by some methods), fractions 2 and 3 are used for cell counts and microbiological examinations, i.e. Gram stain and culture, and fraction 4 is used for the assay of specific proteins and oligoclonal bands and, if required, spectrophotometry (in which case the tube must be protected from light from the moment of collection). Blood for glucose and, in the case of spectrophotometry, for total protein and bilirubin should be taken at the same time.

Appearance

Cerebrospinal fluid is normally crystal clear and colourless. The fluid appears turbid when there are more than 200×10^6 white cells/L ($200/\text{mm}^3$) or 400×10^6 red cells/L ($400/\text{mm}^3$). The presence of bacteria or contamination by epidural fat may also cause turbidity of the specimen. Clot formation may occur when protein concentrations are markedly elevated; where the source of the protein is blood then at least $1\,000\,000 \times 10^6$ red cells/L ($10^6/\text{mm}^3$) are required for clotting to occur. Cerebrospinal fluid may be coloured yellow by bilirubin or, rarely, by carotenoids, red (or more usually pink/orange) by oxyhaemoglobin or brown by methaemoglobin. Yellow discolouration of CSF is called *xanthochromia*, a term that has often been used to encompass the colours imparted by both bilirubin and oxyhaemoglobin. Because the presence of oxyhaemoglobin and bilirubin is most appropriately detected by spectrophotometry, where this technique is used, the term *xanthochromia* should be avoided to minimize confusion.

Cells

The total leukocyte count in normal adult CSF is rarely $>5 \times 10^6/\text{L}$ mononuclear cells (lymphocytes and monocytes). The presence of even one polymorphonuclear leukocyte (neutrophil, polymorph) in the CSF should be a cause for concern in an adult. In neonates, however, the normal cell count is $<30 \times 10^6/\text{L}$ (predominantly polymorphs), with $<10 \times 10^6/\text{L}$ lymphocytes. Changes that occur in the presence of CNS infections form an important part of the initial diagnosis and will be considered later.

Erythrocytes are not normally present in CSF. They appear either following trauma to blood vessels during LP or after an intracranial bleed. A traumatic LP occurs when the needle damages a blood vessel while passing through the vascular epidural space, thereby introducing blood directly into the lumbar sac, and is estimated to occur in 15–20% of all LPs. Traumatic contamination complicates interpretation in two ways.

- The first is in the investigation of suspected intracranial haemorrhage, especially subarachnoid haemorrhage. After haemorrhage, CSF will show equal staining of all samples, while a traumatic tap will

typically result in CSF that is stained initially, but clears in subsequent fractions; in the past this has been confirmed by performing red blood cell counts on the individual collection bottles. While the individual who performed the LP may often have an opinion on whether or not the puncture was traumatic, both clinical impression and sequential red blood cell counting have been demonstrated to be unreliable. The use of such evidence for substantiating an intracranial bleed should be avoided, and if spectrophotometry for detection of oxyhaemoglobin and bilirubin is unavailable or equivocal, the presence or absence of red cells is used to direct further investigations.

- The second is in the interpretation of the CSF white cell count (WCC). Usually an increase in the number of white cells is an indication of infection, but following a traumatic tap, white cells are more likely to be from the contaminating blood. When a CSF leukocytosis is suspected in spite of a traumatic tap, the predicted white cell count can be compared with that observed, as follows:

$$\text{predicted WCC} = \text{CSF red cell count} \times \left(\frac{\text{blood WCC} / \text{blood}}{\text{red cell count}} \right)$$

A ratio of observed (O) to predicted (P) WCC >1 implies CSF leukocytosis. There is a significant overlap of causes for O:P values between 0.75 and 1.0. However, an O:P ratio of <0.1 is highly predictive in terms of excluding infection.

Glucose

The brain has no significant glycogen store and is therefore wholly dependent upon the blood supply of glucose to satisfy its requirements. Although the brain is only about 5% of the total body mass, it uses 20% of the glucose available from the blood. Cerebrospinal fluid glucose concentrations can only meaningfully be interpreted in relation to the plasma concentration, measured on a sample ideally taken within 15 min of LP. Cerebrospinal fluid glucose is derived solely from plasma glucose and is normally 60–80% of the concentration in the latter, although, for the first six months of life, the CSF glucose concentration may equal that of plasma. The CSF:plasma glucose ratio falls below 0.6 in a number of conditions, and this finding can be used as a diagnostic aid. The most notable reductions are observed in bacterial, tuberculous and some fungal meningitides, and in hypoxia, whereas ratios are usually normal in viral meningitis. The exact cause of the lowered CSF:plasma glucose ratio is still unresolved and will be discussed below. Whatever the cause, a lowered ratio usually indicates a diffuse generalized meningeal disease. A false negative result may be obtained if the patient has been given prior treatment with antibiotics.

Glucose is transferred into the CSF by a specific membrane carrier transport system. In the adult, the usual ratio is maintained up to plasma glucose concentrations of 20 mmol/L. For plasma glucose concentrations higher

than this, the CSF glucose concentration does not rise further, probably due to saturation of the transport system. In severe hyperglycaemia, therefore, the CSF glucose may seem disproportionately low relative to that in the blood. It is important that this is not misinterpreted as evidence for infection. Furthermore, the CSF typically takes 2–4 h to equilibrate fully with a change in blood glucose. In a diabetic patient who has recently taken either insulin or an oral hypoglycaemic agent, it is therefore possible to have the paradoxical finding of a CSF glucose concentration that is higher than that in a paired blood sample.

Lactate

Cerebrospinal fluid lactate concentration is normally <2.5 mmol/L; concentrations are controlled independently of those in arterial blood, indicating that it is a product of metabolism within the CNS. An increased CSF lactate concentration is observed in cerebral hypoxia (for example, following a cerebral infarction) and, most notably, in bacterial meningitis, when it is often associated with a decreased CSF glucose concentration. Elevated CSF concentrations of lactate have been demonstrated in patients with inherited disorders of the mitochondrial electron transport chain involving the pyruvate dehydrogenase complex, which give rise to the mitochondrial myopathies. In those with predominantly neurological symptoms, lactate concentrations may be increased only in the CSF.

Proteins

Two-dimensional protein electrophoresis of CSF reveals 200–300 spots, of which about half correspond to 34 unambiguously identified proteins. Nearly all proteins give rise to multiple spots owing to heterogeneity arising from various combinations of glycosylation, phosphorylation and splice variants. Proteins in CSF can be grouped as follows:

- those arising from plasma proteins that have crossed the blood–CSF barriers
- those synthesized in the brain and secreted into the CSF
- those usually present inside CNS cells that have leaked into the CSF following cell damage. Proteins within this group are generally present in only trace amounts.

Table 34.1 lists the ten proteins with the highest concentrations in CSF and plasma.

Some 80% of the total CSF protein concentration is derived from plasma proteins that have diffused passively across the various blood–CSF barriers. The remainder comprises proteins that are synthesized within the CNS (Fig. 34.1). Some plasma proteins, such as transferrin, undergo receptor-mediated transfer. The cells of the choroid plexus also synthesize transferrin and prealbumin, so the CSF concentration is a function of several factors. In health, two factors influence the CSF concentrations of the individual plasma-derived CSF proteins. One is the degree of permeability of the combined blood–CSF barriers to the protein, which is inversely proportional to the Stokes radius of the protein. The other is the plasma concentration of the protein: in general, the CSF

TABLE 34.1 Comparison of the ten proteins with highest concentrations in CSF and plasma

CSF	Mean concentration (mg/L)	Plasma	Mean concentration (g/L)
Albumin	250	Albumin	45
β-Trace (prostaglandin D-synthase)	25	IgG	10
IgG	20	Fibrinogen	3
Transthyretin	17	Transferrin	3
Transferrin	14	α ₂ -Macroglobulin	2.5
α ₁ -Antitrypsin	8	Apolipoprotein A	2
Apolipoprotein A	6	IgA	2
γ-Trace (cystatin-C)	6	Haptoglobin	1.5
Orosomucoid	3.5	α ₁ -Antitrypsin	1.5
Haemopexin	3	Complement factor C3	1.5

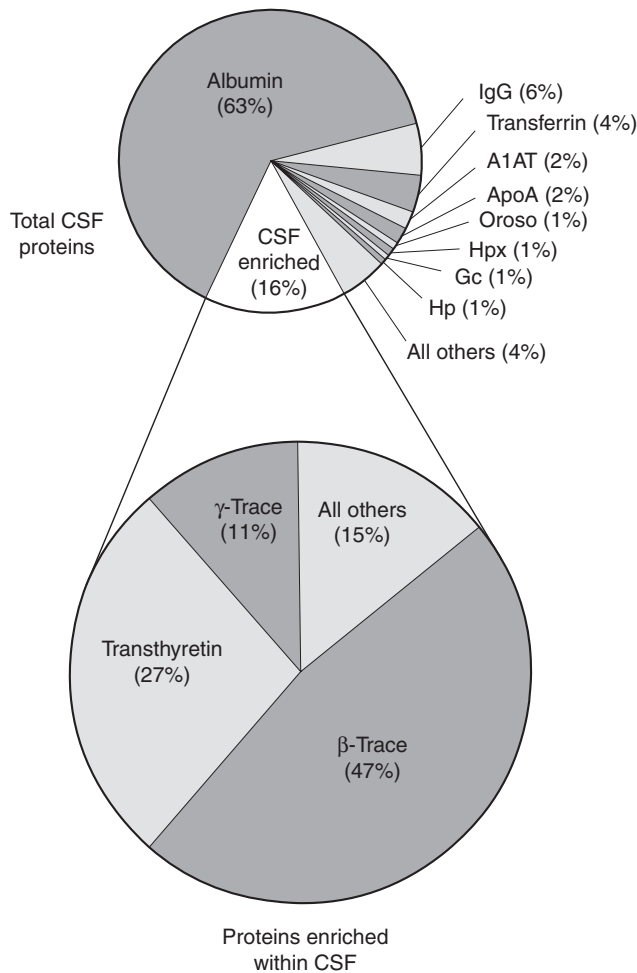


FIGURE 34.1 ■ Protein composition of CSF showing those proteins that are plasma derived and those whose synthesis occurs within the CNS.

concentration is directly proportional to that in plasma, and inversely proportional to molecular size.

All the frequently measured plasma proteins are present in CSF, demonstrating the lack of an upper limit for molecular size exclusion by the blood–CSF barriers. In health, all CSF albumin and immunoglobulin come from the plasma. Albumin accounts for 50–60% of total lumbar CSF protein and is present at a concentration of about 1/250 of that

TABLE 34.2 Age-related values for lumbar CSF total protein, albumin and IgG

Age	CSF total protein (mg/L)	CSF albumin (mg/L)	CSF IgG (mg/L)
Pre-term			
27–32 weeks	1390 (770)	Reliable data not available	Reliable data not available
32–36	1200 (490)		
36–40	930 (290)		
Term			
0–1 weeks	770 (160)	Reliable data not available	Reliable data not available
1–4	660 (160)		
1–3 months	450 (130)		
3–6	290 (40)		
6–12	270 (70)	105 (29)	17 (4)
1–16 years	220 (50)		
17–30	367 (63)		
31–40	363 (61)		
41–50	433 (79)	204 (57)	24 (8)
51–60	479 (94)	242 (56)	27 (9)
61–77	526 (132)	238 (73)	26 (9)

Values are means and (standard deviations).

of the plasma. Permeability of the barrier alters with age; total CSF protein and albumin concentrations reflect this (Table 34.2). The albumin concentration (and thus that of total protein) also increases from ventricle to lumbar regions, reflecting both a change in the permeabilities of the blood–CSF barriers along the CNS axis and the longer time for equilibration of lumbar CSF with plasma.

Cerebrospinal fluid total protein concentrations may increase in two circumstances:

- when there is an increased permeability of the blood–CSF barriers, most commonly owing to infection (e.g. meningitis) or autoimmune inflammatory conditions (e.g. Guillain–Barré syndrome, which is particularly characterized by an acellular CSF and a protein concentration >1 g/L)
- when there is reduced flow of spinal CSF, such as occurs with a partial or complete blockage of CSF flow down the spine (e.g. owing to a prolapsed intervertebral disc, an abscess or a spinal tumour). In this circumstance there is increased equilibration of CSF with plasma, often accompanied by an inflammatory increase in permeability. In Froin syndrome, there is complete spinal block of CSF flow, usually

due to a spinal tumour; cerebrospinal fluid distal to the block stagnates, and this eventually allows complete equilibration of the CSF with plasma leading to protein concentrations that are similar in both fluids.

A comprehensive review of the role of CSF protein analysis in the investigation of CNS disorders is included in the Further reading list.

Assessment of blood–brain barrier permeability and reduced fluid flow

Measurements of total protein, albumin and α_2 -macroglobulin concentrations, and electrophoresis have all been advocated as investigations appropriate to the assessment of barrier permeability. Total protein has been the most widely used and must be interpreted against age-related reference intervals. Some authors have advocated a correction when blood is present due to a traumatic tap, recommending the subtraction of 10 mg/L for every 1000×10^6 erythrocytes/L. The cell count and protein determination must be performed on the same sample of CSF and there should not be any degree of cell lysis, which would otherwise render the correction void. The correction assumes a normal plasma total protein and haematocrit and, while it should be used with caution, provides an approximate measure of the degree to which the CSF protein concentration may be increased following addition of blood from whatever source.

Albumin is a better indicator of barrier dysfunction, for three reasons:

- it is not synthesized by the normal brain, unlike some of the proteins that contribute to CSF total protein
- it is a homogeneous protein, unlike total protein, which is a composite of variable proportions of multiple proteins
- its concentration has a lesser degree of age dependency than that of total protein.

Despite CSF protein concentrations reflecting plasma concentrations, there is generally no discriminatory advantage to be obtained by relating CSF values to those of plasma. The choice between total protein and albumin will depend both on which assay is available at an appropriate analytical standard and cost. Quality assurance programmes indicate similar overall analytical imprecision for total protein and albumin. Experience of using albumin as an alternative to total protein is, as yet, limited.

Intrathecal immunoglobulin synthesis

Intrathecal synthesis of immunoglobulins occurs where B lymphocytes are induced to migrate from the blood into the brain. Once these cells are sequestered in the brain, local cytokine production results in clonal expansion and differentiation into plasma cells, which then start secreting immunoglobulins intrathecally. Intrathecal synthesis of immunoglobulins occurs in a wide variety of neurological diseases, but is most commonly associated with multiple sclerosis and other autoimmune conditions, and infections of the CNS. The immunoglobulin most

widely studied is IgG, and this is associated mainly with subacute and chronic conditions.

Cerebrospinal fluid protein index. The general approach to detecting intrathecal synthesis relies upon the fact that it is possible to calculate the expected CSF concentration for any protein if the plasma concentration of that protein and the integrity of the blood–CSF barriers are known. The integrity of the blood–CSF barrier is given by the CSF/plasma quotient for albumin. If the blood–CSF barriers are intact, the albumin quotient is typically $<7 \times 10^{-3}$; as the barriers become progressively impaired, so the quotient rises, owing to passage of albumin from blood into CSF. Impairment of the barriers will also cause the protein in question to leak from plasma to CSF – but the CSF/plasma quotient for that protein remains roughly in proportion to that of albumin, assuming there is no local synthesis. If local synthesis of protein occurs, then the protein quotient will be increased relative to that of albumin. Thus the ratio of protein quotient to albumin quotient will be increased when local synthesis occurs.

The ratio of quotients is known as the protein index and is given by:

$$\text{protein index} = \frac{\text{CSF protein}}{\text{serum protein}} \times \frac{\text{serum albumin}}{\text{CSF albumin}}$$

This method has been widely used in the past to evaluate intrathecal IgG synthesis, but has now been superseded by examining CSF for IgG oligoclonal bands.

Oligoclonal bands

Locally synthesized IgG demonstrates qualitative differences from that which is normally present. Following serum electrophoresis, the normal gamma globulin region is revealed as a diffuse smear. As CSF gamma-globulins are normally derived from the plasma, they show a similar pattern on electrophoresis. When local synthesis is present, however, electrophoresis of CSF shows multiple discrete bands that are superimposed on the diffuse IgG background. These bands are not seen in the serum. The bands are thought to represent the products of the limited number of clones of plasma cells that are present in the CNS. The pattern is referred to as oligoclonal IgG banding. Due to its higher resolution, isoelectric focusing (IEF) is the method of choice for detecting oligoclonal bands (Fig. 34.2). Oligoclonal bands of IgM and free κ and λ light chains may be detected in a manner analogous to that of IgG.

Papers discussing the relative merits of these approaches (qualitative versus quantitative) are provided in the Further reading list. All the evidence indicates that detection of oligoclonal bands is significantly more sensitive than measuring the IgG index in detecting intrathecal synthesis. The reason is simple: owing to population variation and the compounded imprecision of the four immunoassays necessary to calculate the IgG index, the CSF IgG concentration needs to increase substantially before the index becomes abnormal. Oligoclonal bands, by contrast, can be detected when individual bands

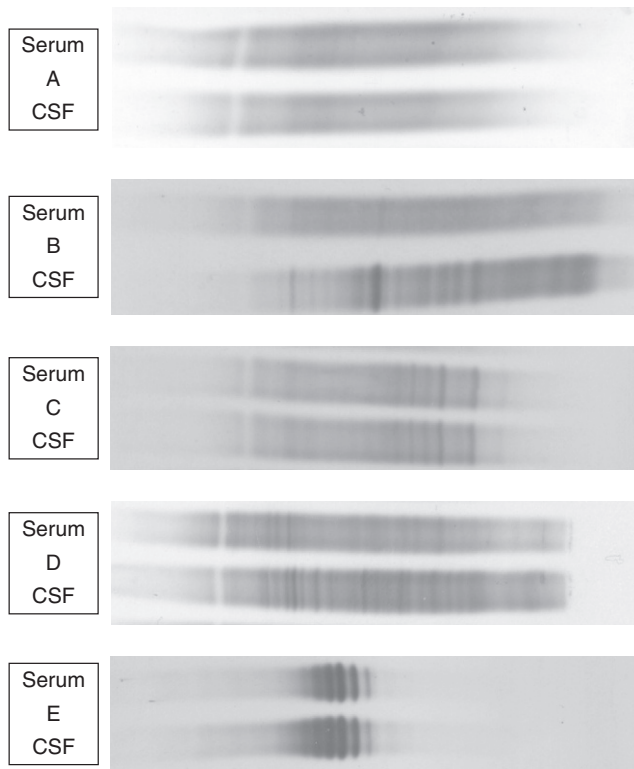


FIGURE 34.2 ■ Isoelectric focusing (IEF) of serum and CSF showing the five basic patterns. (A) normal polyclonal pattern in both CSF and serum. (B) oligoclonal IgG in CSF, absent from serum, typical of local (intrathecal) IgG synthesis. (C) oligoclonal bands in both CSF and serum with an identical band distribution (spectrum), this is termed a 'mirror' pattern and is indicative of a systemic, but not intrathecal, limited clonal IgG synthesis. (D) here, there are oligoclonal bands in both CSF and serum, but there are additional bands in the CSF which are not present in the serum, indicative of both systemic and additional intrathecal limited clonal IgG synthesis. (E) there are identical bands in both CSF and serum. The pattern shows a striking ladder appearance with an identical spacing between adjacent bands. This is highly characteristic of a monoclonal gammopathy.

constitute only 1–2% of the total IgG concentration in CSF. A serum sample must be analysed concurrently with the CSF so that oligoclonal immunoglobulins present in the CSF may be shown to be unique to the CSF, as opposed to those that have originated in the plasma and diffused across the blood–CSF barriers (see Fig. 34.2).

Brain-specific proteins

Measurements of concentrations of proteins synthesized by cells within the CNS and released into the CSF during inflammation or cell destruction are now firmly established tools in the diagnosis of neurological disease. Often described as brain-specific proteins, they are more correctly referred to as brain-enriched proteins. They include proteins such as glial fibrillary acid protein (GFAP), a product of astrocytes; S100, a product of glial cells; 14-3-3, a neuronal cytoplasmic protein; tau and phosphorylated tau, products of the neuronal cytoskeleton; and the amyloid components A β 1-40 and A β 1-42.

Glial fibrillary acid protein is a component of the astrocyte cytoskeleton. When astrocytes are damaged or

undergo activation in response to a variety of cytokine messengers, GFAP may be released into the extracellular compartment from where it diffuses into the CSF. A raised CSF GFAP concentration generally indicates astrocyte damage and/or reactive astrocytosis.

All cells in the body differentially express the S100 family of proteins, of which 14 are currently known. They are found in the cytoplasm and all undergo conformational changes in response to binding calcium. They also interact with key enzymes in a variety of metabolic pathways. In this way, calcium fluxes within the cell can modulate metabolic activity. In the brain, the S100 β homodimer, produced by microglial cells, is the most important member. A high CSF concentration of S100 β is associated with gliosis. In addition, the monomer S100 β acts as a cytokine and is released by microglial cells in response to inflammatory mediators. It is of note that in high concentrations, S100 β can induce apoptosis in cells. As the CSF concentration represents a global average of the production throughout the brain, there may be regional high concentrations causing cell death, which may not always be detected by examining lumbar CSF.

14-3-3 Proteins derive their name from a classification scheme based upon separation by a combination of chromatography and the coordinates of the original protein spot on two-dimensional gel electrophoresis. Once again there is a family of at least ten members, of which 14-3-3 γ is the most important isoform in the CNS. 14-3-3 Proteins are cytosolic and they interact reversibly with a wide range of ligands. Their main physiological role is to connect the cytoplasmic region of cell-surface receptors with their intracellular effector molecules (usually enzyme precursors). 14-3-3 Proteins are ubiquitous in cells, and can make up as much as 1% of the total intracellular protein concentration. Thus 14-3-3 proteins are useful markers of cell disruption, and 14-3-3 γ is particularly helpful in neurology as it is associated primarily with neurons. 14-3-3 γ is a reliable marker of neuronal cell loss when such loss occurs rapidly and extensively, and high CSF values are found in paraneoplastic neurological syndromes, viral encephalitis and neurodegenerative conditions (such as sporadic Creutzfeldt–Jakob disease, CJD). It is not useful in conditions in which neuronal cell loss occurs slowly but progressively (e.g. in Alzheimer disease), because the rate at which it is released is no faster than that by which it is cleared from the CSF.

Tau protein and phosphorylated tau are components of the neuronal cytoskeleton. The phosphorylation of tau is enzymatic and controls the extent to which the protein is incorporated into the cytoskeleton. Tau can be phosphorylated at multiple sites, and in certain conditions, hyperphosphorylated tau is found. It is not yet clear whether the hyperphosphorylation is the cause or a consequence of the underlying conditions. Increased concentrations of total tau and hyperphosphorylated tau are associated with the formation of the neurofibrillary tangles that are seen in several neurodegenerative conditions, especially Alzheimer disease.

A β 1-40 and A β 1-42 are two proteolytic fragments of protein A β (which stands for amyloid protein β). A β is a

constitutive protein that is expressed by nearly all cells of the CNS. The normal metabolism of A β requires it to undergo highly specific proteolytic cleavage to form products that are highly soluble and readily removed. In some neurodegenerative conditions, however, the proteolysis is aberrant, giving rise to larger fragments that are insoluble, and which accumulate in the form of an extracellular amyloid. A β 1-40 and A β 1-42 are two of the normal fragments containing 40 and 42 amino acids, respectively. In some dementias, including Alzheimer disease, there is a *decrease* in the CSF concentration of both A β 1-40 and A β 1-42, which reflects the fact that these components, rather than being in the soluble form (and hence measurable in the CSF), are present as insoluble fragments that are deposited within the CNS.

Cerebrospinal fluid oto- and rhinorrhoea

The leakage of CSF from the intracranial cavity can potentially be serious and usually occurs following skull fractures, although it can be a complication of intracranial surgery, infection or neoplasia, and can also occur spontaneously. Cerebrospinal fluid is characterized by the presence of a β_2 -transferrin band. This is often termed *tautransferrin* but, as there is a risk of confusion with the other tau proteins (see above), the use of the name *asialotransferrin* is preferred. Asialotransferrin is a desialylated form of transferrin. In the blood, desialylated transferrin is rapidly removed (half-life 20 min) by the asialoglycoprotein receptor that is present on the cells of the liver and reticuloendothelial system. The asialoglycoprotein receptor is not expressed in the CNS, so asialotransferrin accumulates in CSF. In normal CSF, asialotransferrin constitutes about 25% of the total transferrin. Asialotransferrin is not present to any great extent in plasma, nasal secretions, tears, saliva or other potentially confounding fluids. Asialotransferrin is easily detected by immunofixation or immunoblotting following electrophoresis (Fig. 34.3). Other proteins that have been used include transthyretin (prealbumin), β -trace (prostaglandin D-synthase) and γ -trace (cystatin-C)

proteins. These are not CSF specific, detectable amounts of all occurring in normal plasma, so interpretation of fluid concentrations is difficult. Measurements of both glucose and chloride concentrations have been shown to lack the discriminatory power of asialotransferrin and should be discontinued.

Haem pigments and ferritin

Oxyhaemoglobin and bilirubin appear in the CSF in a time-dependent manner following an intracranial bleed. Oxyhaemoglobin is released during *in vivo* cell lysis, appearing within a few hours. Following a split into haem and globin, haem is then metabolized to biliverdin by an inducible haem oxygenase followed by reduction to bilirubin by biliverdin reductase; both enzymes are present in cells of the CNS. While animal studies indicate that enzyme induction is maximal by 12 h, this will not necessarily equate to the certainty that an increase in CSF bilirubin will be observable in all patients 12 h after a bleed, given the inevitable biological variation that occurs and possible interspecies differences. Evidence from one apparently definitive study, that CSF bilirubin concentrations were increased by 12 h in all of 111 patients after a proven intracranial bleed, has to be viewed with caution owing to the ambiguous definition of xanthochromia used in this study. Oxyhaemoglobin is also released as a result of *in vitro* lysis of red cells. While initial studies indicated that the degree of lysis is proportional to both red cell count and time delay before separation of supernatant from cells by centrifugation, more recent studies have indicated that the *in vitro* formation of oxyhaemoglobin cannot be reliably predicted. The very low concentration of bilirubin normally present in CSF arises from the passage of albumin-bound bilirubin from plasma into CSF and, hence, will be increased if either or both of the plasma albumin-bound bilirubin and CSF albumin is increased. Under certain circumstances, allowance can be made for this when determining if an intracranial bleed has occurred.

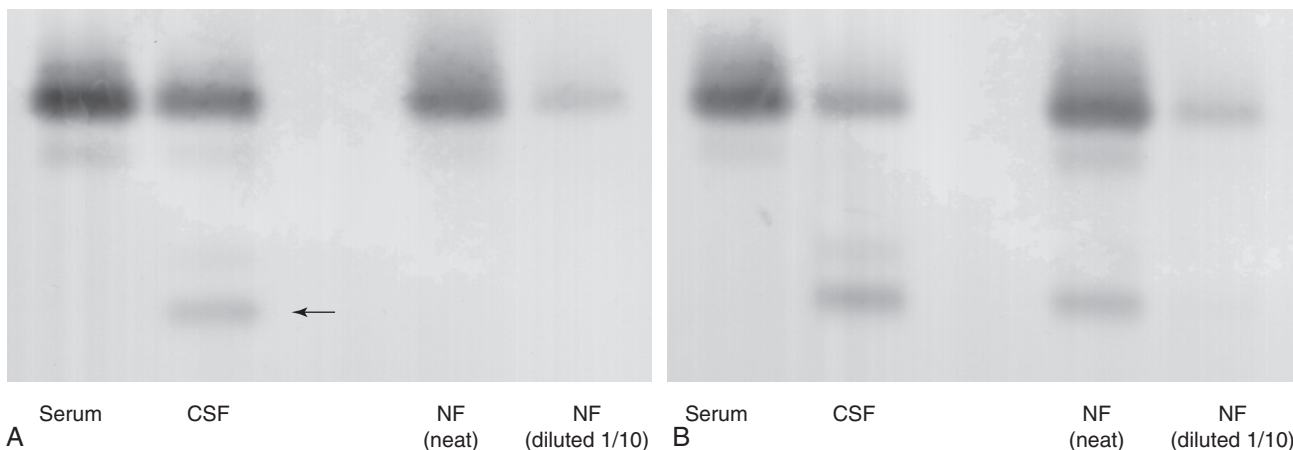


FIGURE 34.3 ■ Transferrin glycoforms in the detection of CSF rhinorrhoea. (A) lane 1, control serum; lane 2, control CSF; lanes 3 and 4, a sample of nasal fluid (NF) which does not contain CSF. The asialotransferrin band is indicated by the arrow. (B) lane 1, control serum; lane 2, control CSF; lanes 3 and 4, nasal fluid containing CSF. In each case, the nasal fluid samples were applied undiluted and diluted 1/10.

Release of iron from haem also stimulates CNS synthesis of ferritin, with subsequent passage of ferritin into the CSF.

Examination of CSF for haem and bilirubin

The currently recommended method for detecting haem and bilirubin in CSF is spectrophotometry, which should cover the range 350–600 nm. An accumulating body of evidence now indicates that visual inspection of CSF for xanthochromia is an unreliable method of confirming an intracranial bleed, because it is insensitive, subjective, non-quantitative, does not distinguish between oxyhaemoglobin and bilirubin and provides no objective record of detection. Direct chemical methods of measuring bilirubin concentration in CSF need to be evaluated rigorously, as does the measurement of CSF ferritin concentration in the investigation of intracranial bleeds.

Enzymes in CSF

Measuring enzyme activities in CSF used to be popular. For example, increased lactate dehydrogenase (LDH) activity has been reported to be a more sensitive and earlier marker than CSF glucose concentration in detecting bacterial meningitis. Unfortunately, however, there is significant overlap between cases diagnosed with bacterial and aseptic/viral meningitis. In addition, false elevations of LDH activity occur in erythrocyte-contaminated CSF. Lactate dehydrogenase activity is also increased in the presence of various primary CNS tumours and when there is meningeal metastatic spread, when the increase involves predominantly the LDH4 and LDH5 isoenzymes.

β -Glucuronidase, creatine kinase BB (CKBB), neuron-specific enolase (NSE), glucose 6-phosphate isomerase (G6PI) and alanine aminotransferase (ALT) are other intracellular enzymes known to be released into CSF by various primary and metastatic tumours. Creatine kinase BB and NSE are also increased during cerebral ischaemia and various meningoencephalitides, thereby limiting their specificities. Ischaemic brain damage can either be reversible or permanent. Although patients suffering major ischaemic events tend to have higher CSF activities of CKBB, LDH, NSE and ALT, the discrimination between recoverable and permanent damage usually requires serial sampling of CSF, and even then the predictive value is low.

Angiotensinogen-converting-enzyme (ACE) is a putative marker of neurological involvement in sarcoidosis. Care must be taken to correct for serum ACE activity, which may be increased in systemic sarcoidosis, and to correct for blood-CSF barrier leakage using either albumin or CSF total protein. The sensitivity of the test is low and the diagnosis of neurosarcoidosis, in the absence of positive CNS histology, remains a presumptive one based on the evidence of CNS inflammation, systemic sarcoidosis and exclusion of alternative diagnoses.

In recent times, measurement of enzyme activities in CSF has generally given way to quantitative immunoassays that measure the enzymes in terms of protein mass. The use of antibodies also means that isoform-specific

immunoassays, which offer significant advantages in terms of determining the tissue of origin over the more crude measurement of overall enzyme activity, can be developed.

Markers of inflammation

Neopterin is the oxidized form of D-7,8-dihydroneopterin, which is produced by macrophages when stimulated by γ -interferon released by activated T cells. Increased concentrations in body fluids are observed following viral, bacterial and fungal infections, as well as other cell-mediated immunopathologies, such as coeliac disease. Cerebrospinal fluid neopterin concentrations are normally very low, as pterins do not readily cross the blood-brain barrier. Increased neopterin concentrations in the CSF thus reflect the presence of activated microglia, the CNS equivalents of macrophages.

There is interest in the measurement of CSF concentrations of neopterin (and dihydroneopterin, which may be the dominant form in CSF, but is more difficult to measure) as a non-specific marker of CNS cellular immune activity. Particular applications have included monitoring the response of multiple sclerosis patients to treatment, detecting possible CNS infections in immunocompromised subjects, such as those with acquired immune deficiency syndrome (AIDS), and establishing evidence of infections when the aetiological agents are difficult to isolate.

β_2 -Microglobulin (β_2M) is an 11 800 Da protein that is non-covalently bound to the class I molecules of the major histocompatibility complex (MHC). While expressed on the surface of virtually all nucleated cells, it is present in highest concentrations on activated lymphocytes and macrophages. Cerebrospinal fluid concentrations of β_2M increase with age and appear not to be correlated with plasma concentrations in diseases affecting the CNS. Increased concentrations of CSF β_2M occur in conditions associated with the presence of activated lymphocytes within the CNS or with increased proliferation and turnover. Measurement of CSF β_2M concentration has been shown to be of value in the early detection of metastatic lymphomas and the spread of leukaemias to the CNS. While both neopterin and β_2M are potentially valuable markers of CNS inflammatory responses, their use as markers of specific pathologies is likely to be limited by the non-specific nature of the inflammatory response.

C-reactive protein (CRP) is a protein synthesized by hepatocytes. Synthesis is increased by several cytokines, but principally interleukin-6, which is released from macrophages that become activated during inflammatory events; this is most notable in bacterial infections, and occurs only to a much smaller degree in viral infections. The inflammatory cells themselves do not synthesize CRP. As with other plasma-derived proteins, the concentration of CRP in the CSF roughly parallels that in the plasma, and is normally $<50 \mu\text{g/L}$. Increased serum concentrations are observed in patients with bacterial meningitis. Measuring CSF CRP concentrations in bacterial meningitis is of limited value, however, as they are highly

variable and are often not as high as might be expected. This is most likely due to CRP binding by bacteria.

Non-biochemical investigations

The investigation of CNS disorders involves many disciplines (Box 34.1). While the naked eye examination of the CSF is a simple test to perform, it is vital to appreciate that it is subjective and unreliable, and that significant abnormalities may be present when the CSF appears to be clear and colourless. The diagnosis of bacterial meningitis is a

BOX 34.1 Non-biochemical investigations of CNS disorders

- Gross appearance of CSF
- Microscopic examination of CSF cell count and differential
- Identification of organisms
 - Gram stain on spun deposit
 - Culture
 - Detection of antigens
 - Demonstration of antibodies
 - Amplification of genetic material
- Imaging
 - Computerized tomography
 - Magnetic resonance imaging
 - Positron emission tomography
- Electrophysiological measurements
 - Visual evoked potentials
 - Auditory evoked potentials
 - Nerve conduction studies

medical emergency and prompt treatment may mean the difference between complete recovery and severe morbidity or even death. Infections cause characteristic changes in the differential white cell count, and sometimes in a Gram stain of a spun deposit (Table 34.3). It is worth noting, however, that white blood cells do not appear in the CSF of patients with bacterial meningitis until ~12 h after the bacteria. These investigations, together with simple biochemical tests (glucose, protein and, if possible, serum CRP and CSF lactate concentrations) should be available as 'stat' tests offering results within 1 h of sampling. Most bacterial infections can be confirmed by standard culture techniques within 24 h. Antigen tests for a number of infective agents are also in use. A particular problem is the positive and rapid identification of viral and fungal infections. With most microbiological infections, polymerase chain reaction (PCR) is the analytical technique of choice in the acute stage of infection. In the subacute and chronic stages, however, PCR tests may be negative. In these instances, recourse to detection of specific antibody responses is necessary. Newer techniques include the detection of antigen-specific oligoclonal bands.

Modern imaging techniques, for example spiral computerized tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET), are important in the diagnosis of many CNS lesions, in conjunction with laboratory investigations of CSF. In some instances, for example intracranial tumours, imaging has largely replaced the need for CSF cytology. Nevertheless, CSF cytology still plays an important role in investigating raised white cell counts in inflammatory conditions and also in detecting and monitoring

TABLE 34.3 Characteristic CSF changes in meningitis and the differential diagnosis

Condition	Appearance	Cells $\times 10^6/L$ (per mm^3)	Gram stain or antigen tests or culture for pyogenic bacteria	Protein (g/L)	Glucose (mmol/L)	Differential diagnosis
Normal	Clear	0–5 lymphocytes ^a	Negative results	0.15–0.4 ^c	2.2–3.3 ^d	–
'Purulent' meningitis	Turbid	100–2000 polymorphs ^b	Positive results	0.5–3.0	0–2.2	Bacterial meningitis ^e Amoebic meningitis Cerebral abscess
'Aseptic' meningitis	Clear or slightly turbid	15–500 lymphocytes (polymorphs may predominate in the acute stage)	Negative results	0.5–1.0	Normal	Viral meningitis ^e Partially antibiotic-treated bacterial meningitis ^f Leptospirosis meningitis Encephalitis Brain abscess TB/fungal meningitis
Tuberculous meningitis	Clear or slightly turbid; fibrin web may develop	30–500 lymphocytes plus polymorphs	Negative results (scanty acid-fast bacilli may be seen in a Ziehl–Neelsen-stained smear)	1.0–6.0	0–2.2	TB meningitis ^e Brain abscess Cryptococcal meningitis

^aIn a neonate, up to 30×10^6 cells/L, mainly polymorphs. ^bA few cases of pyogenic meningitis may have $5–100 \times 10^6/L$ polymorphs. ^cIn neonates, protein concentration up to 1.5 g/L. ^dApproximately 60% of blood glucose concentration. ^eMost frequent cause. ^fGlucose concentration may be reduced.

CNS invasion by leukaemias. Background information on neuroradiological methods and clinical neurophysiological techniques may be obtained from the Further reading list.

BIOCHEMICAL INVESTIGATIONS IN CNS DISORDERS

Acute infections

Infections of the CNS that can present acutely include viral meningitis, bacterial meningitis, cerebral abscess, subdural empyema and viral encephalitis. Acute bacterial meningitis is a life-threatening infection that needs urgent treatment, whereas viral meningitis is generally a less severe disease that is usually followed by complete recovery.

The cellular and microbiological examination of CSF plays a central role in the investigation of suspected infection, and a full description can be found in the Further reading list. Nevertheless, there are occasions when this does not provide a definitive answer; biochemical tests are useful as adjuncts in the diagnosis of infection and need to be available on an urgent basis. Two tests widely employed have been measurement of CSF protein and glucose concentrations, and these have for many years been considered to be essential in the investigation of cases of suspected meningitis. Protein concentration is usually raised, more so in bacterial than in viral meningitis, but there is considerable overlap; rarely, it may be normal. The presence of protein is simply a marker of barrier impairment. Glucose concentrations are usually depressed in bacterial meningitis but often remain normal in viral meningitis. However, CSF glucose concentration may be reduced in some non-bacterial causes of meningitis. Concentrations <1.0 mmol/L are usually found only in severe bacterial meningitis. A CSF:plasma glucose concentration ratio of <0.4 is typical of bacterial meningitis, although a normal value does not exclude the diagnosis.

Cerebrospinal fluid lactate concentrations are raised in bacterial meningitis, but are usually normal (<2.5 mmol/L) in viral meningitis. Concentrations >3.5 mmol/L have been considered indicative of bacterial meningitis, with values <3 mmol/L being more consistent with a viral aetiology. Rapid measurement of CSF lactate has proved of particular value in cases of diagnostic difficulty, for example partially treated bacterial meningitis (when Gram stain and culture are negative and cell counts may be atypical), and early in some viral meningitides when there may be a predominance of polymorphs and variably depressed glucose concentrations. With a threshold of 3.5 mmol/L, the sensitivity and specificity of CSF lactate are greater than those of either CSF glucose or protein concentration, in terms of discriminating between bacterial and viral meningitis.

Measurement of serum CRP concentration may be of value in patients presenting with equivocal CSF findings, with concentrations >100 mg/L being consistent with a bacterial infection. Serum procalcitonin may provide even better discrimination between bacterial and viral

infections. Measurement of serum CRP concentration also enables a rational decision to be made on the duration of antibiotic therapy in patients with brain abscesses or subdural empyemata.

The most reliable indicator of acute infection remains the CSF cell count and differential, with measurement of CSF lactate concentration being used for diagnostically difficult cases. Measurement of cerebrospinal fluid protein appears to add nothing to the diagnosis. Chloride estimation, once widely advocated for the diagnosis of tuberculous meningitis, has been shown to be of no value.

Chronic infections

Chronic infection of the CNS may be caused by viruses, bacteria, spirochaetes, parasites and fungi. In patients who are immunocompromised, most commonly due to cytotoxic drug therapy for cancer, fungal infections and reactivation of latent viral infections are particularly troublesome. In all of these situations, the interpretation of standard PCR techniques can be difficult, as the mere presence of viral or bacterial nucleic acid alone does not always represent a causal relationship. Quantitative PCR, which measures the amount of nucleic acid and so gives an indication of the numbers of organisms present, is superior but still not definitive.

Serological tests for the presence of antibodies are based on the premise that the continued presence of the causative organisms will maintain production of specific antibodies. These tests can suffer from memory effects – the immune system generally maintains a vast memory bank of B cells that manufacture antibodies to individual target antigens. The immune system is frequently upregulated in many inflammatory conditions and this can lead to increased production of antibodies by some or many of these memory cells. If the upregulation occurs in response to a particular infectious agent, then the antibody production will be more specific towards that organism. Antibody production can be measured by enzyme-linked immunosorbent assay (ELISA) or similar technology. However, this approach can make it difficult to distinguish antibodies that vary in affinity from those that vary in quantity, for example a small amount of high-affinity antibody can give the same signal in an ELISA as a larger amount of low-affinity antibody. An alternative approach is to use immunoblotting after isoelectric focusing. This relies on coating a nitrocellulose or similar membrane with the target microorganism. This is then used to immunoblot the CSF and serum following isoelectric focusing. The membrane is subsequently tested for bound immunoglobulin. In this way, oligoclonal bands against the organism in question can be identified. The pattern can then be compared with that found in standard isoelectric focusing (which reveals total IgG oligoclonal bands). If the spectrum of bands seen on the antigen-specific blot resembles that on the total IgG blot, then the causative organism is identified. If, however, the antigen-specific pattern bears no resemblance to the total IgG pattern, then it is unlikely that the organism under study is causally related to the disease. Although the antigen-immunoblotting technique can be extremely

powerful, it is still under development and should be used as an adjunct to classic serological assays.

Haemorrhage and obstruction

Intracranial haemorrhage can occur due to spontaneous rupture of cerebral blood vessels, as a result of injury, malignancy-associated necrosis or as a consequence of bacterial meningitis leading to a diffuse vascular haemorrhage. All may result in the passage of blood into the CSF, with consequent formation of oxyhaemoglobin and bilirubin. By far the most important cause of an intracranial bleed is spontaneous rupture of a cerebral aneurysm into the subarachnoid space, termed a subarachnoid haemorrhage (SAH). In the majority of patients, diagnosis presents no problem. However, a number of studies have shown an initial misdiagnosis rate of up to 30%, often in patients presenting in a good clinical condition with no focal neurological signs, leading to a significantly poorer outcome than in those in whom the correct initial diagnosis has been made.

Following SAH, blood is usually visible on a computerized tomography (CT) scan, and this is the test of choice in the initial examination. This will be positive in over 95% of patients within the first 24h, but in only 50% one week after the bleed. It is in those patients in whom the CT scan is negative or not available that CSF should be examined for oxyhaemoglobin and bilirubin, preferably waiting for 12 h after the onset of symptoms prior to puncture to allow for the formation of bilirubin. Wherever possible, CSF should be examined by spectrophotometry with analysis and interpretation provided according to the guidelines cited in the Further reading list.

Hydrocephalus is a common complication of trauma and obstruction and is best investigated by CT scanning. Shunts inserted to drain CSF into the systemic circulation or peritoneal cavity may become obstructed with a subsequent increase in CSF protein concentration, which may provide an indication that obstruction has occurred.

Inherited metabolic diseases

Biochemical examination of CSF is of value in the diagnosis of certain inherited metabolic diseases.

Patients with primary defects of pyruvate metabolism, for example pyruvate dehydrogenase deficiency or mitochondrial disorders, present with a variety of clinical features. Some, when symptomatic, have a systemic metabolic acidosis with elevated blood lactate concentrations; some may have a near-normal hydrogen ion concentration, mild elevations only of blood lactate concentration but elevated CSF lactate concentration. Measurement of CSF lactate may thus be helpful when such defects are suspected, although normal concentrations do not exclude them. Cerebrospinal fluid lactate determination may also be helpful in certain organic acidemias.

Cerebrospinal fluid amino acid measurement is of established value in confirming the diagnosis of glycine encephalopathy (non-ketotic hyperglycinaemia) in neonates and infants. In this disorder, the CSF glycine concentration is high whereas the blood glycine concentration may

be only mildly elevated or within the reference range for age. The most sensitive discriminator is the CSF:plasma glycine molar ratio, which is increased from the normal 0.025–0.1.

In patients with defects in bipterin metabolism, measurement of CSF bipterin concentrations and the concentrations of the biogenic amine metabolites, homovanillic and 5-hydroxyindoleacetic acids, may be of value in confirming the diagnosis.

Malignancy

In space-occupying lesions of the brain that may be germ cell tumours, the initial approach is to measure serum human chorionic gonadotrophin (hCG) and α -fetoprotein (AFP) concentrations. Pointers to a possible germ cell tumour are a space-occupying lesion in the pineal or suprasellar regions in a male under 15 years of age. The nature of the tumour is usually defined from tissue obtained during surgery. Where operation is contraindicated and serum hCG and AFP concentrations provide no information as to tumour type, there may be a need to measure hCG and AFP concentrations in the CSF.

Rarely, there may be the requirement to examine CSF for monoclonal IgG, IgM, κ or λ when a CNS lymphoma is suspected.

Dementia

Dementia is defined as a global impairment or loss of intellectual function. There are over 50 identified causes of dementia, including various neurodegenerative diseases, such as Alzheimer disease and Parkinson disease; cerebrovascular diseases; infectious diseases; demyelinating disorders; metabolic disorders, such as Wilson disease and the leukodystrophies; malignancies; epilepsies; traumatic brain disease; systemic disorders, such as endocrine disease and vitamin deficiencies, and toxic disorders, including alcohol dependency and drug toxicity. Mental illnesses, such as depression and schizophrenia, may also present with features of dementia. With the increasing life expectancy in developed nations, neurodegenerative diseases, such as Alzheimer disease, are becoming major social burdens. As several types of dementia are treatable, it is important that a diagnosis can be made reliably. Many of these diagnoses are relatively straightforward and require laboratory blood tests that are covered elsewhere in this book. Few involve CSF analysis.

Unfortunately, there are no specific CSF tests for any of the dementias. The differential diagnosis is generally made using a combination of clinical features, CSF findings and the exclusion of other diagnoses. Recent developments in neurogenetics are particularly helpful where there is a familial predisposition, but may not be useful in the majority of cases, which are sporadic in nature.

Cerebrospinal fluid analysis in dementia is generally intended to exclude an infective or demyelinating disorder, as described elsewhere. In specific cases, CSF analysis is intended to detect evidence of neurodegeneration – the death of neurons. Examples include the neuronal

cytoplasmic proteins 14-3-3, neuron-specific enolase (NSE) and neurofilament proteins. Unfortunately, while the intracellular amounts of these proteins are quite high, for example 14-3-3 protein constitutes ~1% of the total cytoplasmic protein of neurons, CSF concentration is generally only raised when there is extensive neuronal loss occurring over a short time period. They can be useful markers in conditions such as sporadic Creutzfeldt-Jakob disease (CJD), paraneoplastic neurological disease and chronic viral infections such as subacute sclerosing panencephalitis (SSPE), but generally they are of little use in detecting slowly progressive neurodegeneration conditions, such as Alzheimer disease or Parkinson disease. In (new) variant CJD, 14-3-3 is a less sensitive marker than in sporadic CJD. The best markers currently available for Alzheimer disease are a combination of A β 1-40 and total tau (see above, Brain-specific proteins). Some dementias are associated with gliosis or astrocytosis, for example Pick disease, and this may lead to an increase in CSF concentrations of S100 β protein and/or glial fibrillary acidic protein (GFAP).

Cerebrovascular disease is second only to Alzheimer disease as a cause of dementia. The usual pathology is subcortical infarcts secondary to embolism or disease of small blood vessels (microangiopathy). Unfortunately, there are no reliable CSF diagnostic tests, although a comprehensive examination of blood haemostatic factors including platelet count, specific clotting factors and antiphospholipid antibodies may be helpful. A strong clinical clue is that cerebrovascular dementia, in contrast to other types, tends to be of acute onset and usually progresses in a stepwise manner.

Cerebrospinal fluid analysis in demyelinating diseases

There are many diseases that affect the myelin sheaths of neurons. One of the most significant is multiple sclerosis (MS), which affects ~125 per 100 000 young adults in the UK, with an annual incidence of ~5 per 100 000. The condition affects twice as many women as men, and usually first presents between the ages of 20 and 40. Autoimmunity is involved in the pathogenesis, but whether as a primary or secondary event is uncertain. Multiple sclerosis is a progressive disorder, typically presenting with a relapsing remitting clinical picture (relapsing remitting MS, RRMS). Some patients have a relentless progressive form of the disease, so-called primary progressive MS (PPMS), while the overwhelming majority of patients who start with RRMS eventually develop a progressive form of the condition, and these are classed as having secondary progressive MS (SPMS). This clinical stratification is important as a guide to choosing an appropriate therapeutic strategy.

The acute lesion of MS is the plaque, which is characteristically a demarcated area of demyelination associated with swelling of axonal cylinders and infiltration by inflammatory cells. This is followed by gliosis and disappearance of the inflammatory infiltrate leading to the formation of a chronic plaque, largely devoid of myelin. Some remyelination can occur during remissions, but this is generally incomplete. The cause of MS is unknown,

but epidemiological evidence suggests an environmental influence, possibly coupled to a genetic predisposition. There is a higher concordance of MS in monozygotic twins compared with dizygotic twins.

Clinical diagnosis of MS can be extremely difficult, particularly in the early stages of the disease when only an isolated focal region of the brain may be affected (this is sometimes termed clinically isolated syndrome). As more of the CNS becomes affected and the symptoms reflect this, then clinical diagnosis becomes easier. The diagnosis depends upon demonstrated clinical evidence for demyelination affecting different areas of the brain at different times. This spatial and temporal separation of lesions is the hallmark of RRMS, and clearly distinguishes MS from similar disorders of myelin, such as acute disseminated encephalomyelitis, in which the clinical presentation of symptoms is typically monophasic. Multiple sclerosis can also be distinguished on MRI by the characteristic individual lesions that are scattered throughout the CNS. Unfortunately, the brain accumulates small areas of damage as part of normal ageing, and it can be difficult to differentiate the lesions of MS from those of age.

While current guidelines, published in 2001 and reviewed in 2005, admit a more prominent role for MRI in the diagnosis of MS than previously, biochemically, the hallmark of MS remains the presence of oligoclonal IgG bands in the CSF that are absent from paired serum. These locally synthesized oligoclonal bands are found in at least 98% of patients with MS and are so indicative of MS that they are incorporated into the pathological definitions of the disease: a clinical diagnosis of MS, based in part on MRI findings has to be called into doubt if there are no oligoclonal bands in the CSF. Oligoclonal bands are not, however, specific for MS and can be found in many other inflammatory and autoimmune conditions affecting the brain. The presence of oligoclonal bands and a CSF protein of >1 g/L points to a diagnosis other than MS. The current gold standard method for detecting oligoclonal IgG is isoelectric focusing coupled to nitrocellulose immunoblotting. This approach is significantly more sensitive at detecting abnormalities of CSF IgG than any of the quantitative measurements (see p. 664).

Additional supportive evidence for MS may be provided by evoked potentials. These are electrical signals generated in response to sensory stimuli. Lesions affecting the optic nerve, the most common isolated site, may be detected by visual evoked potentials, which demonstrate attenuation of amplitude and prolonged latency.

CONCLUSION

Measurement of CSF lactate and glucose concentrations plays a limited but important role in the differential diagnosis of suspected acute CNS infection, particularly when the immediate microbiological examination of the CSF sediment does not provide a diagnosis.

Determination of CSF protein concentration is of limited value in the detection and differential diagnosis

of acute infection. Indeed, it may be misleading, as some have used a normal result as an indication not to proceed with a search for organisms. Protein analysis is of value in suspected Guillain-Barré syndrome and MS, when a value >1 g/L in the presence of oligoclonal bands points to an alternative diagnosis.

In most other disorders of the CNS, routine estimation of protein (and glucose) concentrations, although relatively cheap and apparently simple, does not appear justifiable unless it would truly influence patient management. This rarely appears to be the case where modern imaging techniques are available. The remainder of the tests described have a role appropriate to particular diagnostic problems.

The detection of locally synthesized oligoclonal IgG bands remains crucial in the investigation of suspected demyelinating disease, such as multiple sclerosis. Polymerase chain reaction (PCR) with an appropriate detection system is now the technique of choice in diagnosing acute viral infections, but antibody assays still have a role in chronic infections. At the same time, with increasing emphasis on ameliorating the harmful effects of inflammation, there may emerge a role for the measurement of concentrations of cytokines and other components of the inflammatory response. Continuing research into the various dementias has highlighted several useful biochemical markers that provide diagnostic information.

ACKNOWLEDGEMENT

We would like to acknowledge the contribution of Robert Beetham, who was a co-author of this chapter in previous editions of this book.

Further reading

- Abbott NJ. Evidence for bulk flow of brain interstitial fluid: significance for physiology and pathology. *Neurochem Int* 2004;45:545–52.
A brief overview on the important area of brain interstitial fluid.
- Bauer J, Rauschka H, Lassmann H. Inflammation in the nervous system: the human perspective. *Glia* 2001;36:235–43.
A review of current understanding of the interplay of inflammatory mechanisms involved in neurological disease.
- Beetham R. The examination of cerebrospinal fluid in CT negative suspected subarachnoid haemorrhage. *CPD Bulletin Clinical Biochemistry* 2006;7:25–31.
A critical account of the investigation of CSF in CT negative suspected subarachnoid haemorrhage.
- Cunha BA. The diagnostic usefulness of cerebrospinal fluid lactic acid levels in central nervous system infections. *Clin Infect Dis* 2004;39:1260–1.
A text that emphasizes the usefulness of CSF lactate.
- Graham RNJ, Perriss RW, Scarsbrook AF. Subspecialty Web review – neuroradiology. *Clin Radiol* 2005;60:838–9.
A short review and synopsis of internet-based radiological educational resources that cater for medical students and trainee or non-specialist radiologists, focusing on neuroradiology.
- Huy NT, Thao NT, Diep DT et al. Cerebrospinal fluid lactate concentration to distinguish bacterial from aseptic meningitis: a systematic review and meta-analysis. *Crit Care* 2010;14:R240.
A review summarizing the value of CSF lactate in CNS infections.
- Jerrard DA, Hanna JR, Schindelheim GL. Cerebrospinal fluid. *J Emerg Med* 2001;21:171–8.
A concise review of general laboratory analysis of CSF.
- Keir G, Luxton RW, Thompson EJ. Isoelectric focusing of cerebrospinal fluid immunoglobulin G: an annotated update. *Ann Clin Biochem* 1990;27:436–43.
One of the best critiques available of the preferred method for performing this essential investigation in suspected multiple sclerosis and other diseases involving a humoral response within the CNS.
- McDonald WI, Compston A, Edan G et al. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. *Ann Neurol* 2001;50:121–7.
The guidelines to follow for the clinical investigation of suspected MS. They have recently been revised by Polman et al. (see below)
- Oberascher G. A modern concept of cerebrospinal fluid diagnosis in oto- and rhinorrhoea. *Rhinology* 1988;26:89–103.
This is a small but important area. The paper provides both background and an up-to-date approach to the analysis of watery discharges that might be from CSF leakage.
- Polman CH, Reingold SC, Banwell B et al. Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald Criteria. *Ann Neurol* 2011;69:292–302.
This clarifies the role of spinal cord lesions, and how lesions spread with time.
- Simon L, Gauvin F, Amre DK et al. Serum procalcitonin and C-reactive protein levels as markers of bacterial infection: a systematic review and metaanalysis. *Clin Infect Dis* 2004;39:206–17.
This article reviews the present evidence for use of procalcitonin in distinguishing between viral and bacterial infections.
- Thompson EJ, Keir G. Laboratory investigation of cerebrospinal fluid proteins. *Ann Clin Biochem* 1990;27:425–35.
An overview of the clinically important CSF proteins including brief physiology, pathophysiology and the application of measurement to disease.
- UK National External Quality Assessment Scheme for Immunochemistry Working Group. Revised national guidelines for analysis of cerebrospinal fluid for bilirubin in suspected subarachnoid haemorrhage. *Ann Clin Biochem* 2008;45:238–44.
Current details of how to perform and interpret spectrophotometric scans of CSF.
- Wright BL, Lai JT, Sinclair AJ. Cerebrospinal fluid and lumbar puncture: a practical review. *J Neurol* 2012;259:1530–45.
A pragmatic review on performing lumbar puncture, particularly useful for non-neurologists.

Biochemical aspects of psychiatric disorders

William J. Marshall • Teifion Davies

CHAPTER OUTLINE

INTRODUCTION: PSYCHIATRY AS A CLINICAL DISCIPLINE 673

Investigations in psychiatry 674

THE CLASSIFICATION OF PSYCHIATRIC DISORDERS 674

THE AETIOLOGY OF PSYCHIATRIC DISORDERS 675

BIOCHEMICAL INVESTIGATIONS IN PSYCHIATRIC DISORDERS 675

PSYCHIATRIC MANIFESTATIONS OF ORGANIC DISEASE 676

Acute confusional state (delirium) 676

Anxiety 676

Dementia 677

Depression 678

Post-traumatic stress disorder 679

Schizophrenia 679

ENDOCRINE AND METABOLIC MANIFESTATIONS OF PSYCHIATRIC DISEASE 680

Abnormalities of the hypothalamo–pituitary–adrenal axis 680

Abnormalities of the hypothalamo–pituitary–thyroid axis 680

Abnormalities of the hypothalamo–pituitary–gonadal axis 680

Abnormalities of growth hormone secretion 681

Abnormalities of prolactin secretion 681

Other metabolic abnormalities 681

METABOLIC COMPLICATIONS OF PSYCHOTROPIC DRUGS 681

Lithium 681

Drugs causing hyperprolactinaemia 681

Drugs causing hyponatraemia 681

Drugs causing hyperglycaemia and hyperlipidaemia 682

Drugs interfering with hepatic function 682

FUTURE DEVELOPMENTS 682

CONCLUSION 682

INTRODUCTION: PSYCHIATRY AS A CLINICAL DISCIPLINE

Psychiatry is the branch of medicine that deals with the disturbance, distress and disability arising from those disorders of the nervous system that affect mental functioning. ‘Mental’ functions are those that we regard as distinguishing human beings as persons, and are usually divided into the major domains of cognition (thinking, remembering, planning); perception (awareness of self and environment); mood (feelings, emotions, ‘affect’) and behaviour (objective manifestations of subjective states). Although it is clear that each of these mental functions must be *realized* in the human brain (and that they may be affected by overt damage to the brain, as in head injury), they have been regarded traditionally as *emergent* phenomena that are only tenuously dependent upon the biological structures and processes of the brain itself.

This ‘epiphenomenal’ view of mental functions has been reinforced by two separate sets of findings. First, the psychological and neurobiological study of normal mental functions is only beginning to link them to anatomically discrete structures or pathways. Instead, they have been found to rely on distributed systems of activity in many brain areas that appear to interact in a seemingly limitless pattern of complexity. Second, the delineation of abnormal states has been constrained by a failure to identify specific causes, biological markers or even reliable pathognomonic features, with the result that psychiatry has been slow to follow most other medical disciplines away from the description of clinical syndromes (clusters of clinical symptoms and signs) towards a classification of disorders based on aetiology.

Psychiatric diagnosis remains, therefore, near one extreme of a continuum from ‘organic’ (having often localized pathology, discrete (possibly pathognomonic) signs and objective laboratory findings) to

'non-organic' (unclear structural pathology, overlapping signs and no confirmatory tests). Psychiatrists' skills are largely clinical, and consist in eliciting a wide array of information directly from the patient by means of standardized interviewing and observation of behaviours and responses, supplemented by collateral information from informants such as carers, family members and others. Psychiatric decision-making begins with a comparison of this clinical information with what is known of the normal repertoire of mental states (taking into account age, sex and developmental and cultural factors), and proceeds with a systematization of the abnormal features to produce a set of recognized, but potentially overlapping, clusters. These form hypotheses to be tested by further enquiry, again usually direct questioning, so that the relative likelihood that the patient's presentation is due to one syndrome rather than another is evaluated. The end result of this iterative process is rarely a single clear-cut diagnosis but rather a shortlist of differential diagnoses stratified according to the balance of probabilities on the evidence available. Immediate management will proceed on the basis of the most likely (or most serious in the short term) diagnosis, and will be reassessed iteratively as above and modified as new information, such as response to initial treatment, emerges.

Investigations in psychiatry

The emphasis on clinical information, that is, information gained directly from the patient in the clinic or at the bedside, should not be taken to mean that there is no role for laboratory investigations in psychiatry. Also, a general physical examination should form part of every psychiatric assessment and this or particular features of the patient's history may lead to consideration of appropriate objective tests. What is true of psychiatry, as of all branches of medicine, is that no investigation should be regarded as 'routine': all investigations carry a finite risk of morbidity, and it is important to remember that this can be psychological as well as physical. Therefore, all investigations must be justified and the reasons for their selection or omission should be recorded (see below). It is also of great importance that the rationale for investigation, the intended procedures and potential unwanted effects be discussed with the patient and that his or her consent be sought before proceeding.

The reasons for selecting specific tests will be similar to other branches of medicine: screening (e.g. asymptomatic patients from high-risk population groups); baseline (to exclude or establish the extent of physical disorder, or as a preparation for certain treatments); monitoring (to monitor progression of primary physical disease or of disease secondary to the mental disorder or to monitor therapeutic drugs), and therapeutic (as part of the treatment plan, or to monitor compliance with treatment). A rational plan of investigation should be based on one or more specific purposes, as listed above, and be stratified into: primary (simpler, quicker, cheaper, more likely to gain information); secondary (more complex, expensive or specific) and tertiary (most complex, expensive or reliant on specialist operation).

This stratified approach clarifies the medicolegal status of investigations, at least in psychiatric practice: primary level tests should be considered for every patient, and if not performed, the reasons for omission should be recorded; secondary and tertiary level tests should be performed only if indicated by the presentation, by other findings or on specialist advice, and the reasons for their performance recorded.

THE CLASSIFICATION OF PSYCHIATRIC DISORDERS

Psychiatric illnesses are usually referred to as 'disorders' rather than 'diseases', in part because of the overlap in clinical features between individual conditions and, indeed, with what may be regarded as an extreme of normal behaviour.

There are two principal systems for classifying psychiatric disorders. One is provided by the *Diagnostic and Statistical Manual* (DSM) of the American Psychiatric Association. The current edition (DSM-5), published in 2013, includes 18 groups of disorders, including personality disorders, together with a group of conditions that are deemed to require further research before they can be accepted for clinical use.

The World Health Organization's (WHO) International Classification of Diseases (10th edition) (ICD-10) is more widely used outside the USA. The ICD-10 classification of mental and behavioural disorders is shown in [Table 35.1](#). The number of psychiatric disorders that are recognized is continuing to increase as new conditions are described and others are subdivided and reclassified. The number of psychiatric diagnoses listed in DSM-II in 1968 was 182, and this had slightly more than doubled to 365 in DSM-IV (1994). Although the number appears not to have been significantly increased in DSM-5, several conditions have been sub-divided and others unified and there is an effective overall increase.

Psychiatric disorders have traditionally been divided into neuroses (in which the symptoms, principally anxiety, vary only in severity from those of normal behaviour) and psychoses, with features such as delusions (an abnormal belief, e.g. of persecution) or hallucinations (perceptions lacking an objective stimulus) and typically with little insight. However, this is not to suggest that neuroses are less harmful in their effect. There are also conditions that have features that overlap these definitions (e.g. anorexia nervosa). The classification of some mental disorders as neuroses therefore has little practical value.

Psychiatric disorders are common: the overall prevalence in the general population is estimated to exceed 20%, comprising mainly depression, anxiety and adjustment disorders (e.g. grief reactions) and contribute to about 30% of consultations with family doctors; in hospitals, organic disorders (e.g. delirium, particularly in the elderly) are more common. The major psychoses (by any definition) (e.g. schizophrenia) are less common (<5% in total), but form a major part of the work of many psychiatrists.

TABLE 35.1 A classification of psychiatric disorders

Classification	Examples or notes
Organic disorders	Conditions with a structural (e.g. dementia) or functional (e.g. delirium in fever) basis
Mental and behavioural disorders due to psychoactive substance use	Includes alcohol and illicit or prescribed drug misuse
Schizophrenia and delusional disorders	Delusions are firmly held but abnormal beliefs; hallucinations are clear perceptions without an objective stimulus
Mood (affective) disorders	Depression, bipolar disorder (including mania) etc.
Neurotic, stress-related and somatoform disorders	Physical and/or psychological manifestations of anxiety are central. Includes conditions that suggest pathology in an organ or organs, but where none can be demonstrated
Behavioural syndromes	Having a psychological, physical or physiological cause
Disorders of personality and behaviour	Developmental disorders that do not have an obvious psychiatric or organic cause
Mental retardation	

There are several subspecialties within psychiatry. These include child psychiatry, old age psychiatry, substance misuse and forensic psychiatry.

This chapter primarily concerns the clinical biochemical aspects of psychiatric disorders and those general medical disorders that have psychiatric manifestations.

THE AETIOLOGY OF PSYCHIATRIC DISORDERS

The causation of most psychiatric disorders is multifactorial. Biological (e.g. genetic, organic), psychological and behavioural (e.g. abuse in childhood, emotional trauma), and social and environmental factors (e.g. social isolation) are all involved.

Alterations in neurotransmission are undoubtedly a key mechanism in schizophrenia and affective disorders, although the cause of this remains uncertain. In schizophrenia, for example, there is considerable evidence to implicate dopaminergic pathways, with increased activity in the subcortical and limbic regions of the brain and reduced activity in the prefrontal cortical regions. The drugs that are used in the treatment of schizophrenia all have effects on dopaminergic neurotransmission. Decreased monoamine activity has long been considered to be an important mechanism in depression, a notion that is supported by the efficacy of monoamine oxidase inhibitors in its management. Many addictive drugs have effects on neurotransmission: for example, cocaine blocks the reuptake of dopamine in the brain and benzodiazepines bind to receptors for γ -aminobutyric acid, an important inhibitory neurotransmitter.

Until recently, it has been difficult to study biochemical activity in the brain. Measurements of neurotransmitters and their metabolites made in peripheral blood, blood in veins draining the brain and in cerebrospinal fluid provide only indirect information about brain activity and none on localization. However, the techniques of positron emission tomography (PET) and single-proton emission tomography (SPET) are beginning to shed light on the molecular basis of psychiatric disorders. Functional MRI measures cerebral blood flow, which is

related to neuronal activity. And there can be little doubt that studies on the molecular genetics of receptors and enzymes involved in neurotransmission will contribute to a greater understanding of the biochemical disturbances that underlie psychiatric disorders.

BIOCHEMICAL INVESTIGATIONS IN PSYCHIATRIC DISORDERS

Although laboratory investigations do not have a specific role in the diagnosis of the majority of psychiatric disorders, they are important for several reasons. First, because many physical diseases can give rise to symptoms that occur in psychiatric disorders (e.g. psychosis and delirium in systemic lupus erythematosus), it is frequently important to exclude an organic cause in a patient presenting with an apparent psychiatric disorder (albeit the results of such investigations are usually negative). Features suggesting an organic cause for an apparent psychiatric disorder include late age of onset, with no previous history or family history of psychiatric disorder, and no psychological or social precipitating factor. Second, organic illnesses may be complicated by psychiatric disorders. For example, panic disorder, generalized anxiety, social phobia and depression, occur more frequently in patients with irritable bowel syndrome (IBS) than in the general population, and antidepressant and anxiolytic medication has been shown to benefit significant numbers of patients with refractory IBS. Another connection between organic illness and psychiatric disorders is that the latter may cause both metabolic and endocrine disturbances (e.g. amenorrhoea in anorexia nervosa).

Third, psychotropic medication can cause metabolic abnormalities, and particularly given the high prevalence of psychiatric disorders, and hence the use of psychotropic drugs, clinical biochemists must be cognizant of these, so that, for example, the finding of a high serum prolactin concentration in a patient being treated with an antipsychotic drug should not lead automatically to a search for a pituitary tumour. Fourth, substance abuse may involve the clinical biochemist both analytically and in relation to the metabolic disturbances that may ensue. While it is beyond the scope of this book to discuss the

TABLE 35.2 Laboratory investigations to exclude organic disease in patients with psychiatric symptoms

Investigation	Rationale
Full blood count	Anaemia can lead to cerebral hypoxia, causing confusion; macrocytosis may be due to vitamin B ₁₂ deficiency, a cause of dementia, and to a high alcohol intake; chronic anaemia may contribute to depression
Acute phase proteins	High erythrocyte sedimentation rate (ESR) suggests systemic illness, e.g. infection, malignancy, that may cause delirium; C-reactive protein (CRP) elevation may indicate abrupt onset of inflammation
Creatinine, 'electrolytes', estimated glomerular filtration rate (eGFR), calcium, liver function tests	Electrolyte disorders and renal and hepatic failure can cause delirium; they may result from behavioural disturbance secondary to psychiatric disorder
Blood gases	Hypocapnia can mimic anxiety (but note that, more frequently, hyperventilation secondary to anxiety or a panic attack can cause hypocapnia); hypercapnia in respiratory failure can cause delirium
Glucose	Hypoglycaemia can mimic anxiety (and, when chronic, may lead to behavioural disturbance and trigger referral to a psychiatrist); hyperglycaemia can cause delirium
Thyroid function tests	See text
Syphilis serology	General paresis (a late manifestation of neurosyphilis), presenting with dementia, tremor and upper motor neuron signs, can develop 5–15 years after primary infection, if this is not treated
Blood cultures	To detect occult sepsis
Blood alcohol	A frequent cause of and exacerbating feature in psychiatric presentations

Other laboratory investigations that may be valuable (though are less frequently required) include measurement of heavy metals, carboxyhaemoglobin (for carbon monoxide poisoning), antinuclear factor (for cerebral lupus erythematosus) etc.

role of the laboratory in the detection of substance abuse, the metabolic complications associated with some of the more frequently encountered substances are discussed in Chapter 40.

Table 35.2 summarizes the biochemical and other laboratory investigations that may be of value in excluding an organic cause for an apparent 'psychiatric presentation'; the ensuing sections of this chapter describe the major psychiatric manifestations of organic diseases and the metabolic changes that can occur in psychiatric disorders and as a result of their treatment.

PSYCHIATRIC MANIFESTATIONS OF ORGANIC DISEASE

Acute confusional state (delirium)

Delirium is a complex syndrome of altered consciousness, manifest by decreased awareness of the environment and inattention, together with cognitive impairment, with memory deficit, disorientation, decreased perception and language disturbance. There may, in addition, be emotional disturbances (e.g. anxiety, irritability) and psychomotor changes (e.g. agitation or retardation). Symptoms tend to fluctuate, confusion typically being greater at night or in unfamiliar surroundings (e.g. a hospital ward).

The causes include:

- metabolic and endocrine (see below)
- infection (systemic, particularly with high fever and meningeal/cerebral)
- vascular (e.g. cerebral haemorrhage and infarction)
- toxic (e.g. drugs, both intoxication and withdrawal, and toxins, e.g. carbon monoxide)
- neoplasia (primary cerebral tumours, cerebral metastases and paraneoplastic syndromes)

- trauma (e.g. subdural haematoma, cerebral contusion)
- miscellaneous (e.g. post-ictal, postoperative).

The metabolic and endocrine causes are summarized in Table 35.3. When one of these is the cause of acute confusion, there will often be other clinical features of, or clues to, the condition, but if this is not the case, the laboratory investigations performed beyond a standard 'biochemical profile' and full blood count (with differential) should reflect the relative frequency of the conditions. It should be noted that in the elderly, almost any acute illness can present with acute confusion, and that the elderly may be at increased risk of developing confusion in response to any causative condition as a result of poor nutrition, visual or auditory impairment, or incipient dementia. The tendency for elderly people to be taking multiple therapeutic drugs coupled with age-related changes in the rates of metabolism and excretion of drugs is another important factor.

Sepsis and drug intoxication or withdrawal are particularly common causes of delirium. Sepsis may increase the risk of drug-induced delirium by increasing the permeability of the blood–brain barrier. For drugs that are protein-bound, a low plasma albumin concentration may also increase toxicity by increasing the proportion of the drug present in the unbound form.

Anxiety

Anxiety is a normal response to a threat, but in the anxiety disorders, the response is out of proportion to any real danger and causes suffering and behavioural disturbance.

The anxiety disorders comprise three conditions: generalized anxiety disorder, in which symptoms are persistent, and two conditions in which they are episodic – panic disorder, in which there is no apparent stimulus,

TABLE 35.3 Metabolic and endocrine causes of delirium (acute confusional state)

Cause	Notes
Cerebral hypoxia	Secondary to, for example, cardiac failure, respiratory failure, hypotension, severe anaemia
Hypo-/hyperglycaemia	
Hyponatraemia	Dependent on the rate of fall as well as the absolute value of the sodium concentration
Hypo-/hypercalcaemia	
Renal failure	Depression is a far more frequent psychiatric manifestation of hypothyroidism
Liver failure	
Adrenal failure	
Hyper-/hypothyroidism	
Hyperpyrexia	
Acute porphyria	Causing rhabdomyolysis as in malignant hyperthermia or the 'neuroleptic malignant syndrome' due to central dopamine blockade
Vitamin deficiencies (thiamin, niacin and vitamin B ₁₂)	
Drugs	Including alcohol (ethanol), illicit agents and a wide range of therapeutic drugs, particularly in the elderly

and phobic anxiety disorders, in which anxiety stems from fear of situations, for example going out alone or being in crowded spaces (agoraphobia), or objects, for example spiders. All types can be associated with somatic symptoms, including chest pain, palpitation, dyspnoea (typically a feeling of not being able to take a full breath), paraesthesiae and sweating, but these are particularly associated with panic disorder. They may lead to the patient seeking urgent medical advice because they fear a serious illness, for example a heart attack. Patients with generalized anxiety disorder may seek medical advice because of bowel disturbance.

The somatic symptoms in the acute anxiety disorders are due to a combination of increased sympathetic activity, release of catecholamines and hyperventilation. The latter causes falls in arterial PCO_2 and hydrogen ion concentration (rise in pH) with normal bicarbonate and normal PO_2 , reducing the plasma concentration of ionized calcium and increasing neuromuscular excitability. Typically, symptoms can be provoked by over-breathing and abated by rebreathing into a paper (not plastic, which risks asphyxiation) bag. It is not usually difficult to distinguish hyperventilation in anxiety from hyperventilation in response to acidosis (e.g. ketoacidosis) or in pulmonary disease (e.g. asthma, pulmonary oedema) but, if necessary, arterial blood gases can be measured.

Hypocapnia can also cause anxiety. The symptoms produced (e.g. chest pain) may lead sufferers to believe that they have serious disease, inducing anxiety and setting up a vicious circle.

The differential diagnosis of anxiety disorders includes drug dependence and withdrawal, hyperthyroidism, hypoglycaemia, hypoparathyroidism and pheochromocytoma.

Thyroid function tests should always be performed in patients presenting with an apparent primary anxiety disorder. Anxiety, insomnia, emotional lability and difficulty concentrating are frequently present in patients with hyperthyroidism and the symptoms may occasionally be sufficiently severe to suggest hypomania. It may be difficult to distinguish between an anxiety state and mild hyperthyroidism: features suggesting the latter include eye

signs, the presence of a goitre and proximal myopathy. Tremor is often present in both conditions, but the hands tend to be warm and moist in hyperthyroidism as a result of the hyperdynamic circulation, but cold and clammy in anxiety disorder. Resting pulse rate is usually normal in anxiety disorders but increased in hyperthyroidism. The anxiety symptoms of hyperthyroidism usually respond to treatment with β -blockers.

The diagnosis of hypoglycaemia is discussed in detail in Chapter 17, and of pheochromocytoma in Chapter 38. Measurement of fasting blood glucose concentration during an attack is a simple matter; measuring plasma or urinary catecholamines or their metabolites is not, but pheochromocytoma, although rare, should be considered when there are predominant sympathetic symptoms (e.g. palpitation, sweating) and no obvious precipitating factor can be identified.

Dementia

Dementia – a condition characterized by a progressive decline in cognitive function (particularly loss of short-term memory) without confusion or loss of arousal – is common, particularly in the elderly, in whom Alzheimer disease and cerebrovascular disease (usually diffuse, small vessel disease) are the major causes. Metabolic and endocrine conditions are uncommon causes. A treatable cause should be sought through laboratory investigations and imaging, particularly in younger patients. The former should include a full blood count, erythrocyte sedimentation rate (ESR) and standard biochemical renal- liver- and bone-related tests, thyroid function tests, measurement of serum vitamin B₁₂ concentration and serology for systemic lupus erythematosus and syphilis (see Table 35.2). In individuals with a history of alcohol abuse, thiamin deficiency should be considered but is usually diagnosed on the basis of the therapeutic response to an intravenous bolus of thiamin. Heavy metal and carbon monoxide poisoning may need to be considered if there is possibility of exposure. A syndrome of 'pseudodementia' may arise in severe depression; this should be sought clinically as investigations are likely to be normal.

Considerable effort continues to be invested in research to determine the cause of Alzheimer disease and to identify early markers of the disease that could direct early intervention. To date, however, although various candidate markers have been investigated, none has been identified and the role of laboratory investigations is confined to eliminating treatable causes of dementia.

The pathological hallmarks of Alzheimer disease are senile plaques and neurofibrillary tangles throughout the cerebral cortex but particularly in the hippocampus. The first stage in the development of plaques is the deposition of β -amyloid protein, which is produced through the proteolysis of amyloid precursor protein (APP). This protein is coded for on chromosome 21, which may explain why individuals with Down syndrome (trisomy 21) are prone to the early development of Alzheimer-like changes in the CNS. Neurofibrillary tangles develop as a result of the aggregation of tau protein driven by its undergoing hyperphosphorylation, but the underlying cause of either of these processes remains unknown. As the disease develops, diffuse cerebral atrophy ensues. The most obvious functional abnormality in Alzheimer disease is a reduction in cerebral cholinergic activity. This has led to the development of treatments using cholinesterase inhibitors, but thus far, these are of limited benefit, and that only in early disease.

Most cases of Alzheimer disease are sporadic, but a small number show autosomal dominant inheritance, and are due to mutations in the genes coding for APP or presenilins 1 and 2. Genetic factors also impinge on sporadic Alzheimer disease: some 80% of patients possess at least one $\epsilon 4$ allele of the gene coding for apolipoprotein E (the usual allele being $\epsilon 3$). However, the basis of this association remains uncertain.

Depression

Introduction

Depression is one of the commonest psychiatric disorders. Its typical features are low mood, loss of interest and enjoyment (anhedonia), reduced energy and increased fatigue; other common features include psychomotor retardation, poverty of movement and disordered ideation (e.g. hypochondriasis, feelings of unworthiness). Somatic symptoms, particularly headache and fatigue, are common. In severe cases, auditory hallucinations, delusions and suicidal ideation may be present. Recurrent distinct episodes are common, although some patients experience a chronic, low-grade depression (dysthymia). Depression can occur on its own (unipolar disorder) or (less frequently) be a feature of bipolar affective disorder (manic depressive psychosis) with episodes of depression or hypomania interspersed with periods of normality.

Depression can develop as a natural response to physical illness (depressive adjustment disorder) but is typically not as severe as depression arising without precipitating factors. If the associated illness is self-limiting, it may only be transient. However, the boundary between depressive adjustment disorder and depressive illness is not clear-cut, and depressive symptoms arising in a patient with physical illness should not be assumed to be a

response to the illness. Indeed, because some depressive symptoms can be features of organic illness, the possibility of depression should be actively considered. Up to a third of patients with physical illness have depressive symptoms, and major depression is present in up to 15%, the prevalence correlating with the number of physical symptoms. Untreated comorbid depression worsens the outcome in physical illness (e.g. cardiovascular disease, diabetes, arthritis). Indicators of possible depressive disorder include: failure to adjust to the illness; reduced physical functioning; slower recovery than might be expected, and a reduction in social activity.

Depression and thyroid function

The differential diagnosis includes several other psychiatric disorders, and depression can coexist with other disorders, for example panic disorder. The most frequent organic disease that may be mistaken for depression is hypothyroidism, although apathy is more common than true depressive symptoms in the latter. Other features include inattentiveness and loss of short-term memory. Frank psychosis due to hypothyroidism ('myxoedema madness') is now rare. Hypothyroidism is a common condition, particularly in the elderly; its manifestations can be protean and it should be excluded by measuring thyroid stimulating hormone (TSH) and free thyroxine concentrations in any patient presenting with depression or a decline in cognitive function. The psychiatric manifestations improve with thyroxine replacement treatment, although sometimes the cognitive dysfunction persists, perhaps because of a deleterious effect of chronic thyroid hormone deficiency on the CNS. Depression that is refractory to treatment may show a pattern of 'sick euthyroidism' (low plasma TSH and free thyroxine concentrations, see Chapter 19), and the depression may respond to thyroxine treatment; thyroid function should be reassessed when the depression improves as there is unlikely to be primary hypothyroidism. In some patients, the initiation of thyroxine treatment can temporarily exacerbate the psychiatric symptoms or even, in susceptible individuals (e.g. those with a family history of affective disorder), precipitate a manic episode. It should be remembered that, particularly in elderly people, hyperthyroidism occasionally presents atypically (apathetic hyperthyroidism) and may be mistaken for depression. Thyroid disease is discussed in detail in Chapter 19.

Depression and adrenal function

Depression is a common feature of Cushing syndrome, whether pituitary dependent or independent or the result or administration of exogenous steroids (though this can also cause euphoria). That it is cortisol, rather than corticotrophin (adrenocorticotrophic hormone, ACTH) that is the cause of the depressive symptoms is suggested by the absence of any excess of depression in patients with Nelson syndrome (in which ACTH concentrations are high but cortisol is low). Furthermore, metyrapone, which inhibits the synthesis of cortisol and causes increased concentrations of ACTH, has a beneficial effect on depression in Cushing syndrome. Depression in

Cushing syndrome is more frequent in women and in older age groups, and its severity shows some correlation with the severity of other features of the condition. The acute administration of high doses of corticosteroids (e.g. prednisolone 30 mg/day) can lead to hyperphagia, insomnia and euphoria, and their sudden withdrawal can precipitate depression.

Patients with depression often fail to show suppression of plasma cortisol concentration in the overnight dexamethasone suppression test (see Chapter 18) and (though less frequently) in the formal low-dose test, in which dexamethasone 0.5 mg is administered 6-hourly for 48 h.

Depression can lead to accelerated loss of bone mineral density and increased risk of fracture: the hypercortisolaemia is likely, in part, to be responsible for this.

Depression can also be a feature of Addison disease. The symptoms may pre-date the classic physical signs of the condition but respond to cortisol replacement treatment.

Depression in the metabolic syndrome and diabetes

There is a well recognized association between diabetes mellitus and depression. Dr Thomas Willis, a 17th century British physician, considered that diabetes was caused by 'sadness or long sorrow'. One study has reported a prevalence of major depression in diabetes of 9.3%, compared with 6.1% in the general population. The presence of depression in diabetes is associated with poorer glycaemic control and poorer outcome but appropriate psychiatric intervention can reduce the morbidity associated with diabetes. Both diabetes and depression are risk factors for cardiovascular disease; the risk in patients with depression is approximately twice that in those without.

It seems more likely that diabetes predisposes to depression than vice versa. A nine-year prospective study of people with or without depression showed no excess risk of diabetes in those with depression. Further, treatment of depression in patients with diabetes has no effect on glycaemic control. How diabetes predisposes to depression is uncertain, but insulin resistance as part of the metabolic response to the psychosocial stress arising from having a chronic illness may, in part, be responsible.

The major underlying pathogenic factor in type 2 diabetes is resistance to the actions of insulin. Insulin resistance also underlies the metabolic syndrome (see Chapter 15). This syndrome comprises a range of abnormalities, including visceral obesity, dyslipidaemia, hyperglycaemia and hypertension that are risk factors for both type 2 diabetes and cardiovascular disease. The prevalence of the metabolic syndrome in patients with diabetes is high. Several other abnormalities have been described in both depression and the metabolic syndrome, including hypercortisolaemia, abnormal autonomic nervous system function (e.g. increased resting heart rate with reduced heart rate variability), endothelial dysfunction, increased platelet reactivity and evidence of persistent inflammation.

Plasma concentrations of nutritional factors such as 25-hydroxyvitamin D and ω -3 fatty acids have been found

to be consistently low in depressed patients, the former being linked to reduced activity and exposure to sunlight, the latter to modulation of cytokine activity leading to increased hypothalamo-pituitary-adrenal and inflammatory activity. There is no evidence of clinical benefit from dietary supplementation.

Post-traumatic stress disorder

This syndrome can arise as a result of exposure to any severe psychological trauma; physical trauma may also be involved. Examples of causes include experience of natural disasters, war, terrorism, witnessing violent death, torture and rape. Its features include:

- recurrent, intrusive recollections of the event ('flashbacks')
- feelings of detachment and isolation (avoidant symptoms)
- persistent symptoms of increased arousal (e.g. exaggerated startle reflex).

The severity of the condition in individuals is dependent both on factors related to the event itself (e.g. its severity, the proximity of the individual to it) and on the personality of the individual, any history of psychiatric illness and other factors.

Endocrine abnormalities associated with the post-traumatic stress disorder include decreased cortisol secretion secondary to decreased secretion of ACTH (despite increased secretion of corticotrophin releasing hormone, CRH). The ACTH response to CRH is reduced, and that to dexamethasone increased, suggesting enhanced negative feedback. This contrasts with the elevated cortisol and reduced sensitivity to dexamethasone that is characteristic of several other psychiatric disorders. The urinary excretion of catecholamines is increased and there is evidence of increased CNS noradrenaline (norepinephrine) and serotonin activity, these two neurotransmitters being known to have a role in storing and retrieval of memory. Some patients with post-traumatic stress disorder have increased plasma concentrations of thyroid hormones and an enhanced TSH response to TRH (thyrotrophin releasing hormone), yet are clinically euthyroid.

Functional magnetic resonance imaging (fMRI) of the brain suggests features in common between post-traumatic stress disorder and obsessive-compulsive disorder.

Schizophrenia

The history of schizophrenia research is littered with the discovery, and subsequent abandonment, of putative biomarkers of the disorder or its aetiology. Recent interest has focused on elevated plasma concentrations of inflammatory cytokines found in patients with both first episode psychosis and chronic schizophrenia. Acute inflammation has been reported to increase the relapse rate of schizophrenic symptoms, and the widespread direct and indirect influences of cytokines on brain neurotransmitter pathways and neural plasticity suggest a mechanism. Augmentation of antipsychotic treatment with the anti-inflammatory drugs acetylsalicylic acid (aspirin) and

celecoxib (a cyclooxygenase-2 inhibitor) has been advocated and may improve treatment response in some patients.

Despite the major functional abnormalities that may be present in patients with schizophrenia, significant metabolic disturbances are relatively uncommon. Basal concentrations of most hormones are normal in schizophrenia (elevations in vasopressin are an exception), although the TSH and gonadotrophin responses to their respective releasing hormones are typically blunted. However, significant metabolic abnormalities may develop as a consequence of treatment with psychotropic drugs, as described later in this chapter. Furthermore, substance abuse is frequent in patients with schizophrenia and this may cause metabolic abnormalities, e.g. hyponatraemia with amfetamines (particularly methylenedioxymethamphetamine, MDMA, 'ecstasy'), hypoglycaemia with ethanol (see Chapter 17) and decreased secretion of gonadotrophins with opioids.

ENDOCRINE AND METABOLIC MANIFESTATIONS OF PSYCHIATRIC DISEASE

The hypothalamus is central to the regulation of the endocrine system and the sympathetic nervous system, and is involved in behavioural responses, so it is not surprising that many psychiatric diseases affect endocrine function. Not all do, however: schizophrenia (see above) is a notable exception.

Abnormalities of the hypothalamo–pituitary–adrenal axis

Abnormalities of cortisol secretion are common in psychiatric disease. Approximately 50% of patients with depression have hypercortisolaemia (though with preservation of the normal circadian variation) secondary to increased secretion of CRH and hence ACTH, although the response of ACTH secretion to an intravenous bolus of CRH is reduced (in contrast to patients with Cushing disease, in whom it is typically exaggerated). Enhanced vasopressin secretion may also contribute to the increased secretion of ACTH in depression.

The reason for the increased CRH drive is unclear. It may be a result of impairment of the normal inhibitory feedback of cortisol on the hypothalamus, but this could be part of the underlying disorder or a response to it.

Individuals with alcohol dependence may develop features of Cushing syndrome (pseudo-Cushing) but these usually resolve rapidly on withdrawal of alcohol.

Similarity between the features of chronic fatigue syndrome and adrenal insufficiency have prompted research into endocrine function in the former condition. Plasma cortisol concentrations tend to be slightly lower than normal, although adrenal responsiveness to ACTH is not impaired, and the ACTH and cortisol responses to insulin-provoked hypoglycaemia are normal. Treatment with low doses (5–10 mg/day) of hydrocortisone has been shown to be of benefit in some patients.

Plasma cortisol concentrations are frequently decreased in patients with post-traumatic stress disorder; these patients often demonstrate increased sensitivity to the suppressive effects of dexamethasone on cortisol secretion. This contrasts with the findings in many other psychiatric disorders, in which hypercortisolaemia is accompanied by reduced sensitivity to dexamethasone.

The ACTH response to CRH is blunted in some patients with anxiety disorders but plasma cortisol concentrations are normal. Hypothalamo–pituitary–adrenal function has been reported to be normal in patients with schizophrenia.

Abnormalities of the hypothalamo–pituitary–thyroid axis

Patients with depression frequently demonstrate abnormalities of thyroid function. Up to 20% of patients with depression have been reported to have 'sick euthyroidism' (see above) or subclinical hypothyroidism, with normal thyroxine concentrations but slightly elevated concentrations of TSH secondary to increased secretion of TRH and with increased TSH responsiveness to TRH. This pattern is more common in patients with bipolar disorder. Other patients (more particularly those with unipolar depression) have a blunted TSH response to TRH. However, this is not specific to depression, having been reported in schizophrenia and alcoholism as well. Treatment with antidepressants leads to normalization of thyroid function tests, although it is of interest that treatment with tri-iodothyronine apparently enhances their efficacy in some patients, both increasing the rate at which they become effective and the overall response. This suggests that there may be a specific link between thyroid function and mood disorder rather than to the symptoms just being a consequence of reduced cerebral metabolism. Tri-iodothyronine is widely used by psychiatrists for this purpose.

Patients with other acute and chronic psychiatric illnesses may demonstrate the features of the sick euthyroid syndrome (see Chapter 19), and patients with anorexia nervosa typically demonstrate the abnormalities characteristic of starvation, with a low plasma thyroxine concentration, increased concentration of reverse tri-iodothyronine and a reduced TSH response to TRH.

Abnormalities of the hypothalamo–pituitary–gonadal axis

The secretion of gonadotrophin releasing hormone (GnRH) is impaired at low body weights (see Chapter 18), and oligo-/amenorrhoea is frequently reported by women with anorexia nervosa. Amenorrhoea is often a feature of moderate to severe depression, and may result from diminution of the release of follicle stimulating hormone (FSH) and luteinizing hormone (LH) from the anterior pituitary. Pulsatile release of both FSH and LH takes place during slow-wave (i.e. non-REM) sleep (SWS), but this phase of sleep is shortened or absent in moderate to severe depression. Testosterone has been reported to have an anti-depressant effect in depressed hypogonadal men.

Abnormalities of growth hormone secretion

Anorexia nervosa typically also causes increases in plasma growth hormone concentration with exaggerated responses to growth hormone releasing hormone (GHRH) but decreased concentrations of insulin-like growth factor 1 (IGF-1), probably due to a combination of weight loss and increased concentrations of glucocorticoids

Two major abnormalities of growth hormone secretion have been described in depression: hypersecretion during daytime and reduced secretion during sleep. As growth hormone is also secreted from the pituitary in bursts during SWS, the latter abnormality may arise from the reduction of SWS that is found in depression.

Abnormalities of prolactin secretion

Anxiety is a common cause of hyperprolactinaemia, and blood for prolactin measurement should always be drawn under conditions that minimize stress.

Nocturnal plasma prolactin concentrations are increased in some patients with schizophrenia, but elevated prolactin concentrations in this condition are usually a consequence of treatment with dopamine receptor antagonists (see below).

Prolactin secretion is usually normal in depressive illness, although responsiveness to stimuli such as L-tryptophan may be diminished.

Other metabolic abnormalities

Several psychiatric disorders can lead to abnormalities of fluid, electrolyte and acid–base balance. The respiratory alkalosis that can occur as a result of hyperventilation in anxiety states has been mentioned above. Abnormalities associated with drug abuse are discussed in relation to individual drugs in Chapter 40. Examples include hypophosphataemia in alcoholism and alcohol withdrawal, and potassium and magnesium depletion in patients abusing diuretics or laxatives. Severe hypokalaemia can develop in anorexia nervosa secondary to chronic starvation; this may be exacerbated by vomiting and laxative abuse. Patients with bulimia are particularly prone to developing potassium depletion with hypochloreaemic alkalosis as a result of self-induced vomiting. Hypophosphataemia may also occur but is particularly associated with refeeding (see p. 211). Multiple deficiencies of mineral nutrients can occur in both conditions, and in depression. Psychogenic polydipsia can cause profound hyponatraemia and may be difficult to distinguish from diabetes insipidus, as the chronically high urine output can lead to ‘medullary washout’ with a secondary decrease in the concentrating capacity of the kidneys (see Chapter 4).

METABOLIC COMPLICATIONS OF PSYCHOTROPIC DRUGS

Lithium

Lithium is a valuable mood stabilizing drug, widely used in the management of bipolar disorder, both for prophylaxis and in the treatment of acute episodes (particularly

of mania). However, it has the potential to cause severe adverse effects. These are directly related to its plasma concentration, and monitoring of plasma concentrations every three months is mandatory (see Chapter 39).

Lithium reduces the responsiveness of the distal parts of the nephron to vasopressin (antidiuretic hormone), resulting in a reduced capacity to concentrate the urine. This effect is demonstrable in most patients being treated with the drug but is rarely of clinical significance, although a few patients develop nephrogenic diabetes insipidus. This is partially reversible if lithium is withdrawn. Lithium is 95% excreted by the kidneys and dehydration decreases the glomerular filtration rate, so increasing the concentration of the drug and potentially exacerbating the toxic effect. However, provided that acute toxicity is avoided, lithium rarely causes long-term renal damage.

Lithium can also interfere with thyroid function and cause primary hypothyroidism, the risk of this being greater if thyroid autoantibodies are present. For these reasons, plasma creatinine (and hence estimated glomerular filtration rate (eGFR), thyroid stimulating hormone (TSH) concentrations should be measured before starting lithium treatment and at regular intervals (e.g. 4–6 monthly) thereafter.

Drugs causing hyperprolactinaemia

Prolactin secretion by the anterior pituitary is inhibited by dopamine, and given the fact that several antipsychotic drugs (e.g. phenothiazines, butyrophenones) are dopamine (D₂ receptor) antagonists, it is not surprising that these drugs are a frequent cause of hyperprolactinaemia. Metoclopramide and domperidone, two widely used antiemetic drugs, also have this effect. Prolactin concentration may be increased to as much as 2500 mU/L, well into the range seen in patients harbouring microprolactinomas. Clinical features of hyperprolactinaemia are, however, relatively uncommon, but amenorrhoea, galactorrhoea and breast enlargement may occur in women, and gynaecomastia and impotence in men. Hyperprolactinaemia is less frequent with the newer second generation antipsychotics, for example olanzapine and quetiapine (though it is seen with risperidone), whose actions are either non-dopaminergic or have higher affinity for 5-hydroxytryptamine (serotonin) receptors than dopamine receptors.

Drugs causing hyponatraemia

Several drugs used in the management of patients with psychiatric disease can interfere with water homeostasis (probably through stimulating vasopressin secretion) and lead to hyponatraemia. They include selective serotonin reuptake inhibitors (e.g. paroxetine) and carbamazepine, primarily an anticonvulsant but also used in the prophylaxis of bipolar disorder. In the majority of patients, any resultant hyponatraemia is only mild and is asymptomatic, but there is a risk of significant water retention leading to intoxication, and serum sodium concentration (and less frequently osmolality) should be monitored in patients being treated with these drugs.

Drugs causing hyperglycaemia and hyperlipidaemia

The prevalence of diabetes mellitus is higher in patients with schizophrenia in comparison with the general population, but the introduction of chlorpromazine in 1956 caused a significant further increase, and the term 'chlorpromazine diabetes' was used to describe this. The mechanism appears to be insulin resistance. Newer second generation antipsychotics can also cause hyperglycaemia and hypertriglyceridaemia, and sometimes hypercholesterolaemia, albeit less frequently. Clozapine is particularly implicated, followed by olanzapine; metabolic abnormalities are less common with risperidone and quetiapine, and least with aripiprazole. Weight gain is frequent with both clozapine and olanzapine. It is recommended that glycated haemoglobin (HbA_{1c}) and lipid concentrations should be measured at inception of treatment with any antipsychotic, and repeated twice yearly in patients on clozapine (they also require to be under surveillance for agranulocytosis, a potentially lethal adverse effect) and olanzapine, and annually for other antipsychotics.

Drugs interfering with hepatic function

Many drugs can cause mild hepatic abnormalities, usually manifest as an increase in plasma aminotransferase and/or alkaline phosphatase activity. Examples include chlorpromazine, haloperidol and some tricyclic antidepressants. Chlorpromazine can cause cholestasis, and this may persist after the drug has been discontinued. The hepatotoxicity of valproate is a particular problem. This drug, primarily used as an anticonvulsant, has become one of the most frequently used mood-stabilising drugs. Several cases of fatal hepatotoxicity have been reported. Patients present with lethargy, anorexia, nausea, and jaundice are found to have hyperammonaemia (although elevated ammonia concentrations have been reported in patients taking valproate in the absence of overt toxicity). Unfortunately, this reaction is an idiosyncratic one, being dependent neither on the dose nor the concentration of the drug, so that therapeutic monitoring of concentrations is of no value in preventing it.

Many drugs used in psychiatric practice affect plasma protein binding (notably carbamazepine) or the hepatic cytochrome enzyme system, causing changes in the availability of other drugs utilizing the same pathway. The result may be to reduce the therapeutic effects of a drug or increase its adverse effects.

FUTURE DEVELOPMENTS

Simple 'mono-neurotransmitter' models of psychiatric disorders are increasingly untenable and, along with them, simple distinctions between mental and physical illness. Evidence is accumulating of the interplay of neuroendocrine and inflammatory processes in many

psychiatric disorders, and that some disorders that cluster epidemiologically (depression and cardiovascular disease) might not be independent but manifestations of the same pathogenetic process. Although it seems unlikely that measurements of individual or even groups of simple analytes in peripheral blood will ever become of value in the diagnosis or management of the major psychiatric disorders, it is possible to envisage an increasing role for the clinical biochemist in psychiatry. Pharmacogenetics – the study of the genetic influence on responsiveness to drugs – has the potential to facilitate a more tailored approach to prescribing, in relation both to the selection and dosage of drugs. And, given that some psychiatric disorders are clearly at least in part of genetic origin, the identification of genes conferring susceptibility to these conditions will enable the development of microarray-based tests to identify individuals at high risk. These could then be targeted for intervention to manage other risk factors, or possibly prophylactic drug treatment.

CONCLUSION

At present, most psychiatric disorders are diagnosed clinically. The purpose of biochemical investigations in patients with psychiatric symptoms is to rule out an organic cause. However, many psychiatric disorders can give rise to metabolic and endocrine abnormalities, as can treatment with psychotropic drugs.

Further reading

De Jonge P, Roest AM. Depression and cardiovascular disease: the end of simple models. *Br J Psychiatry* 2012;201:337–8.

A brief review of the interdependence of physical and psychiatric disorder. Gelder M, Andreasen N, Lopez-Ibor J et al. *New Oxford textbook of psychiatry*. 2nd ed. Oxford: Oxford University Press; 2012.

An advanced single volume textbook of psychiatry.

Gelder M, Mayou R, Geddes J. *Psychiatry*. 3rd ed. Oxford: Oxford University Press; 2005.

A good introduction to the language of psychiatry, the major psychiatric disorders and their management.

Harmer CJ, Cowen PJ, Goodwin GM. Efficacy markers in depression. *J Psychopharmacol* 2011;25:1148–58.

A review of the effects of depression and its treatment on key psychological factors such as emotional processing.

Krishnan V, Nestler EJ. The molecular neurobiology of depression. *Nature* 2008;455:894–902.

An overview of the many neurobiological (behavioural, biochemical, electrophysiological) approaches informing current theories of depression.

Osborn DPJ, Wright CA, Levy G et al. Relative risk of diabetes, dyslipidaemia, hypertension and the metabolic syndrome in people with severe mental illnesses: Systematic review and metaanalysis. *BMC Psychiatry* 2008;8:84. <http://www.biomedcentral.com/1471-244X/8/84>.

An extensive review of the occurrence and features of metabolic syndrome in major psychiatric disorder, with emphasis on the psychoses.

Pariante C, Lightman S. The HPA axis in major depression: classical theories and new developments. *Trends Neurosci* 2008;31:464–8.

A short review of the role of the HPA axis in stress and depression, and the possibility of targeted treatments.

Taylor D, Paton C, Kapur S. *The Maudsley prescribing guidelines in psychiatry*. 11th ed. Chichester: Wiley-Blackwell; 2012.

Comprehensive coverage of the use of drug treatments, their effects and how to monitor them.

Biochemical aspects of neurological disease

Paul Hart • Clare M. Galtrey • Dominic C. Paviour • Min Htut

CHAPTER OUTLINE

INTRODUCTION 683

ENCEPHALOPATHY 683

Toxic and metabolic encephalopathy 685

Septic encephalopathy 687

Autoimmune encephalopathy 688

Dementia 688

SPINAL CORD DISORDERS 688

Vitamin B₁₂ deficiency (subacute combined degeneration of the spinal cord) 688

Folate deficiency 689

Copper deficiency 689

Vitamin E deficiency 689

Hepatic myelopathy 689

Hexosaminidase A deficiency 689

Adrenomyeloneuropathy 689

PERIPHERAL NEUROPATHY 689

Small fibre painful axonal neuropathy 691

Diabetic neuropathies 691

Immune mediated neuropathies 692

Acute inflammatory neuropathies and variants 692

Chronic inflammatory demyelinating polyneuropathies and variants including paraproteinaemic neuropathies 692

Chronic kidney disease and established renal failure 693

Liver disease 693

Endocrine disturbances 693

Nutritional peripheral neuropathies 693

Neuropathy associated with bariatric surgery 694

Strachan syndrome 694

Metabolic neuropathies 694

Mitochondrial disorders 696

Paraneoplastic neuropathies 696

MOVEMENT DISORDERS 697

ATAXIA 699

Friedreich ataxia 699

Ataxia with isolated vitamin E deficiency 699

Abetalipoproteinemia 700

Ataxia telangiectasia 700

Early onset ataxia with oculomotor apraxia and hypoalbuminemia 700

Fragile X-associated tremor/ataxia syndrome 700

Hexosaminidase deficiency (GM2 gangliosidosis) 700

Cerebrotendinous xanthomatosis (cholestanolosis) 700

Neuronal ceroid lipofuscinosis 701

Celiac disease 701

INFLAMMATORY DISORDERS OF THE CENTRAL NERVOUS SYSTEM 701

CONCLUSION 701

INTRODUCTION

Despite the significant advances seen in laboratory and imaging diagnostics, many of the conditions seen by neurologists are still diagnosed on purely clinical grounds. In neurology, there remains no substitute for a meticulous clinical history and careful examination.

Neurological investigations include all imaging modalities, neurophysiological tests of both the central and peripheral nervous system, and laboratory based investigations, including haematology, biochemistry,

immunology, histopathology and cytology. This chapter will focus on the neurological conditions for which biochemical abnormalities are pertinent. Separate chapters cover the biochemistry of cerebrospinal fluid and muscle disease and other conditions that overlap with general neurological practice.

ENCEPHALOPATHY

'Encephalopathy' is a term for any diffuse disease of the brain that alters brain function or structure. It is a common condition that encompasses coma, acute confusional

BOX 36.1 Some causes of metabolic encephalopathy

- Toxic
 - Alcohol
 - Drugs
 - Carbon monoxide
- Nutritional disorders
 - Thiamin deficiency (Wernicke–Korsakoff syndrome)
 - Vitamin B₁₂ deficiency
- Organ failure
 - Hepatic failure
 - Established renal failure
 - Cardiorespiratory failure (hypoxic/ischaemic and hypercapnic)
- Diabetic disorders
 - Hypoglycaemia
 - Hyperglycaemia (diabetic ketoacidosis and hyperosmolar hyperglycaemic state)
- Disturbances of sodium and water balance
 - Hyponatraemia
 - Hypernatraemia
- Hypercalcaemia
- Septic encephalopathy

state, delirium and dementia. The hallmark of encephalopathy is an altered mental state.

Encephalopathy is caused either by processes that directly affect the structure of the brain or by systemic and metabolic factors (see [Box 36.1](#)). Encephalopathy can occur acutely (over hours to days); subacutely (over weeks to months) or chronically (over years) ([Table 36.1](#)). Although the mechanism of many causes of encephalopathy is well understood, the pathophysiology of metabolic encephalopathy is less clear but is likely to involve changes in amino acids and neurotransmitter profile.

For patients presenting with altered consciousness, a detailed history often has to be taken from family, friends and other healthcare professionals, to determine pre-existing medical conditions and time of onset of altered mental state. For example, patients presenting with altered consciousness following apnoea or cardiac arrest are likely to be suffering from a hypoxic–ischaemic encephalopathy but alternative causes need to be excluded. Access to medication, previous psychiatric problems or use of illicit drugs may suggest intoxication. If the patient is known to have diabetes, hypo- or hyperglycaemia is likely. Patients on diuretics are at risk of hyponatraemia.

TABLE 36.1 Neurological causes of encephalopathy

Acute (hours to days)	Trauma Infection (e.g. meningitis/encephalitis) Vascular (e.g. stroke) Space occupying lesion
Subacute (weeks–months)	Infection (e.g. HIV, progressive multifocal leukoencephalopathy, Lyme disease) Neoplastic (primary or secondary) Inflammatory (e.g. cerebral vasculitis) Autoimmune encephalitis (immune-mediated, paraneoplastic, Hashimoto) Degenerative (e.g. Creutzfeldt–Jakob disease; mitochondrial disease)
Chronic (years)	Degenerative (Alzheimer disease, Parkinson disease)

Both a general systemic examination and detailed neurological examination are required to determine the cause of encephalopathy. Specific breath smells suggest certain causes, e.g. ammonia-like (uraemia); fruity-smelling (ketoacidosis); musty or fishy (acute hepatic failure); onion (paraldehyde), and garlic (organophosphates). Hypertension may indicate amphetamine or cocaine intoxication or a hypertensive encephalopathy. Examination of the skin may show signs of chronic liver disease, haemodialysis fistula, previous sternotomy scars from cardiac surgery or injection marks from intravenous drug use.

Neurological examination should be divided into focal (localizing) and global (non-localizing) signs. Global signs suggest a process affecting the whole brain and include altered consciousness (acute confusional state, delirium and coma). The level of coma can be quantified with the Glasgow Coma Score, for which the minimum score is 3/15 ([Table 36.2](#)).

Other global neurological signs include generalized seizures, tremor, asterixis (abnormal jerking tremor in hands seen in liver flap) and myoclonus (involuntary brief jerk-like twitching, often multifocal). Focal signs indicate a localized problem within one part of the brain, for example visual or language disturbance from cerebral cortex pathology or eye movement disturbance, slurred speech or swallowing problems from brainstem pathology. Focal neurological signs usually suggest a primary neurological cause of encephalopathy.

TABLE 36.2 Glasgow coma scale (GCS)

The total is obtained by adding together the scores for each column

Best eye response	Best verbal response	Best motor response
No eye opening	No verbal response	No motor response
Eyes open in response to pain	Incomprehensible sounds	Extension to pain
Eyes open to speech	Inappropriate words	Abnormal flexion to pain
Eyes open spontaneously	Confused	Withdrawal to pain
	Oriented	Localizes to pain
		Obeys commands

Two-thirds of encephalopathies are secondary to systemic metabolic factors. It is therefore important that these are excluded before assuming there is a neurological cause. Many metabolic encephalopathies are reversible if corrected promptly. Investigations for a patient with encephalopathy should include blood tests, urinalysis, imaging studies, and, when indicated, cerebrospinal fluid examination and an electroencephalogram. Increasingly, laboratory investigations can be used to make a definite diagnosis of neurological causes of encephalopathy owing to the improved recognition of antibodies for autoimmune encephalitis syndromes and biomarkers in neurodegenerative disease (Table 36.3).

Toxic and metabolic encephalopathy

Toxic encephalopathy is caused by exogenous substances including solvents, drugs, radiation, paints, industrial chemicals, and certain metals. Some of the more common causes are discussed below. More details can be found in Chapter 40.

Carbon monoxide

This colourless and odourless gas is the most common cause of accidental poisoning in Europe and North America, as well as being a major cause of suicidal deaths. When 20–30% of total haemoglobin is bound to carbon monoxide, it causes headache, nausea and shortness of breath; when 50–60% is bound, coma results. Carbon monoxide exposure is detected by measuring the carboxyhaemoglobin concentration in blood. Carbon monoxide poisoning is treated with oxygen, using hyperbaric facilities in severe cases, if available.

Alcohol

Alcohols, including ethanol, produce an altered mental state by central nervous system depression. Methanol and ethylene glycol similarly cause direct central nervous system depression but are also toxic via metabolism in the liver to formic acid (methanol) and glycolic and oxalic acids (ethylene glycol).

Methanol and ethylene glycol concentrations can be measured in the blood but usually the suspicion is raised by other biochemical abnormalities. Poisoning with either causes a high anion gap metabolic acidosis. Treatment is by blocking the action of alcohol dehydrogenase with ethanol or fomepizole to decrease the production of toxic metabolites.

Opioids

Overdose causes coma, pinpoint pupils and respiratory depression. This can be from iatrogenic, accidental or suicidal overdose (e.g. morphine) or from use of illicit drugs (e.g. heroin). Diagnosis is usually suspected from the clinical triad but a rapid urine screening test can confirm opioid exposure. Treatment is with naloxone, given as a bolus dose followed by an infusion if the patient responds, in addition to supportive care.

TABLE 36.3 Investigations for encephalopathy

Test	Cause of encephalopathy
First-line	
Blood/serum	
Full blood count	Elevated white cell count (infection), macrocytosis (alcohol, hypothyroidism)
Serum electrolytes (osmolality if required)	Hypo- and hypernatraemia, hypercalcaemia
Glucose	Hypo- and hyperglycaemia
Urea and creatinine	Uraemia
Liver function tests	Hepatic failure
Arterial [H ⁺] (pH)	Diabetic ketoacidosis and hypercapnic encephalopathy
Arterial PO ₂ and PCO ₂	Hypoxia/ischaemia and hypercapnia
C-reactive protein	Infectious
Erythrocyte sedimentation rate	Infectious and autoimmune
Thyroid function tests	Hashimoto thyroiditis
Vitamin B ₁₂ and folate	Nutritional deficiency
Urine	
Ketones	Diabetic ketoacidosis
Drug and alcohol screen	Toxic
Other	
CT brain	Structural
Second-line (as appropriate to the suspected clinical condition)	
Blood/serum	
Ammonia	Hepatic failure, inborn errors of metabolism, sodium valproate
Angiotensin-converting-enzyme	Sarcoidosis
Anti-nuclear, anti-neutrophil cytoplasmic, anticardiolipin antibodies, extractable nuclear antigens	Cerebral vasculitis
Thyroid peroxidase antibodies	Hashimoto thyroiditis/steroid responsive encephalopathy
Lactate dehydrogenase	Lymphoma
Syphilis, <i>Borrelia</i> and HIV serology	Infectious
Voltage gated potassium channel, N-methyl-autoimmune D-aspartate receptor, glutamate receptor and glutamic acid decarboxylase antibodies	Paraneoplastic
Antineuronal antibodies	Paraneoplastic
Cerebrospinal fluid	
Microbiology	Infectious, inflammatory, autoimmune and paraneoplastic
Protein, glucose	Infectious
Viral PCR	Inflammatory
Oligoclonal bands	Paraneoplastic
Antineuronal antibodies	Creutzfeldt–Jakob disease
14-3-3 protein	Carcinomatous meningitis
Cytology	
Other	
MRI brain	Structural, infectious, autoimmune
Electroencephalography	Seizures

Thiamin (vitamin B₁) deficiency

Wernicke–Korsakoff encephalopathy is a clinical triad of confusion, ataxia and ophthalmoplegia (eye movement disorder). It comprises two syndromes. First, Wernicke encephalopathy, an acute or subacute confusional state with ataxia and ophthalmoplegia and second, Korsakoff dementia, characterized by severe amnesia and confabulation. The former is often reversible but the latter is irreversible and persistent. Wernicke–Korsakoff encephalopathy is caused by thiamin deficiency and is most commonly seen in alcoholics but can also occur in patients with anorexia, hyperemesis gravidarum, those receiving total parenteral nutrition, or in the context of malabsorption and refeeding syndrome. The role of measurement of plasma thiamin concentrations in diagnosis and monitoring remains controversial and functional assessment by assay of red cell transketolase is no longer used by most laboratories for diagnosis. Therefore, laboratory tests are aimed at excluding other conditions and providing evidence of risk factors (e.g. macrocytosis and deranged liver function tests in alcoholism). Brain magnetic resonance imaging (MRI) can show signal change in the thalami, mammillary bodies, tectal plate and periaqueductal grey matter. Treatment is with parenteral thiamin. In the UK, normal practice is to give thiamin in a vitamin complex of riboflavin, nicotinamide, pyridoxine and ascorbic acid. Early treatment can reverse the symptoms rapidly and prevent dementia. Administration of intravenous glucose (including as part of parenteral nutrition) to patients who are severely malnourished, can exhaust their (already depleted) supply of thiamin and precipitate Wernicke–Korsakoff encephalopathy. Thiamin should therefore be administered before starting a glucose infusion in patients at high risk.

Vitamin B₁₂ deficiency

This causes a typical pattern of degeneration of the white matter producing encephalopathy, myelopathy (subacute combined degeneration of the spinal cord), peripheral neuropathy and optic neuropathy. Vitamin B₁₂ concentrations are easily measured and, if deficiency is corrected early, further damage can be prevented (See p. 688 and p. 693).

Liver failure

Hepatic encephalopathy occurs as a consequence of liver function disturbance and its severity is graded from 0 to 4 (Table 36.4).

Two situations can lead to severe hepatic encephalopathy – acute fulminant hepatic failure and decompensation of chronic liver disease. In the UK, acute liver failure is most commonly from paracetamol overdose or viral hepatitis, and chronic liver disease from alcohol abuse or chronic hepatitis B or C infection. Encephalopathy is thought to develop because hepatocellular dysfunction produces neurotoxins (e.g. ammonia), false neurotransmitters and benzodiazepine-like substances. In chronic liver disease, there can be the additional effect of porto–systemic shunting. This allows significant quantities of ammonia, formed in the bowel from protein, to reach the systemic circulation.

TABLE 36.4 Grading of hepatic encephalopathy (West Haven classification system)

Grading	Symptoms
0	Minimal hepatic encephalopathy. Lack of detectable changes in personality or behaviour. Minimal changes in memory, concentration, intellectual function, and coordination. Asterixis absent
1	Trivial lack of awareness. Shortened attention span. Impaired addition or subtraction. Hypersomnia, insomnia, or inversion of sleep pattern. Euphoria, depression, or irritability. Mild confusion. Slowing of ability to perform mental tasks. Asterixis can be detected
2	Lethargy or apathy. Slurred speech. Obvious asterixis. Drowsiness, lethargy, gross deficits in ability to perform mental tasks. Obvious personality changes, inappropriate behaviour and intermittent disorientation, usually regarding time
3	Somnolent but can be aroused. Unable to perform mental tasks. Disorientation about time and place, marked confusion, amnesia. Occasional fits of rage, present but incomprehensible speech
4	Coma with or without response to painful stimuli

The severity of the neurological disorder correlates poorly with ammonia concentrations so measurement rarely contributes to management. ‘Liver function tests’ will confirm that the concentration of bilirubin is high, often with abnormal liver enzyme activities and, more importantly, poor synthetic function with abnormal clotting. Plasma paracetamol concentration can be measured if overdose is a possibility. Patients with acute liver failure may require liver transplantation for survival but the encephalopathy of chronic disease may be reversed by conservative measures (see Chapter 14 for further details).

Chronic kidney disease and established renal failure

Uraemic encephalopathy presents with apathy, fatigue, inattentiveness and irritability, followed by confusion, hallucinations, slurred speech, tremor and asterixis. The mechanism is unclear but correction of uraemia reverses the encephalopathy. Suggested mechanisms include retention of organic acids, elevation of phosphate concentration in the cerebrospinal fluid (CSF), increased calcium content of the cerebral cortex owing to the action of PTH, and alteration of concentrations of neurotransmitters or proinflammatory cytokines. Although the degree of the rise in urea correlates with the severity of encephalopathy, it is not thought to be causative. The investigation, monitoring and treatment of uraemia are discussed in Chapter 7.

Respiratory failure

Hypercapnia occurs in respiratory failure either secondary to lung disease (e.g. chronic obstructive pulmonary disease) or to mechanical problems such as neurological

disease (e.g. myasthenia gravis). Clinically, hypercapnia presents with headache, papilloedema, mental slowing, drowsiness, confusion, coma and asterixis. The mechanism is unclear but thought to be due to a direct effect of carbon dioxide possibly on the hydrogen ion concentration (pH) of the CSF. Hypercapnia can be confirmed by measurement of PCO_2 on an arterial blood sample. Coma can occur with $PCO_2 > 9$ kPa. Treatment is of the underlying cause and, once corrected, there is no prolonged cerebral damage (see Chapter 5 for further details).

Cardiorespiratory failure

Hypoxic–ischaemic encephalopathy is caused by hypoxaemia and/or reduced blood flow to the brain from failure of the cardiac system, the respiratory system or both (e.g. myocardial infarction, drowning). When cerebral perfusion pressure falls, there is a failure of autoregulation leading to ischaemia and hypoxia and to cell death by both necrosis and apoptosis. The diagnosis is usually evident from the history. Myoclonus is common (Lance–Adams syndrome). There is a wide spectrum of clinical outcomes depending on the degree and duration of the insult from full recovery to severe prolonged disability and death.

Disorders of glucose metabolism

Hypoglycaemia causes confusion, seizures and coma and can cause permanent neurological damage if not reversed quickly. Up to 15% of patients with diabetes will have at least one episode of hypoglycaemic coma in their lifetime. Recurrent hypoglycaemia, such as that caused by an insulinoma, can be mistaken for epilepsy. The brain contains 1–2 g of glucose stored as glycogen, which is utilized within approximately 30 min if blood glucose concentration remains low. Ketones can also be used as an energy source but this is not sufficient in prolonged hypoglycaemia, nor are they available when hypoglycaemia is caused by high insulin concentrations. The most common causes of hypoglycaemia are accidental or deliberate overdose of insulin, insulinoma, and depletion of liver glycogen (e.g. acute liver failure, alcohol binge, severe starvation). Hypoglycaemia is easily detected by blood glucose measurement at the bedside (although it should be confirmed by laboratory measurements) (see Chapter 17 for further details).

Hyperglycaemic coma occurs in two forms – diabetic ketoacidosis and the hyperosmolar hyperglycaemic state. Coma is a late feature of diabetic ketoacidosis and develops when hyperglycaemia, dehydration, acidosis and shock are severe. Cerebral oedema is a rare but often fatal complication both of the condition and its treatment. The coma of the hyperosmolar hyperglycaemic state usually develops more insidiously and occurs in elderly patients, often with previously undiagnosed diabetes mellitus. Extreme hyperglycaemia with greatly increased plasma osmolality, electrolyte losses and dehydration exacerbated by impaired thirst awareness, are all thought to contribute to the development of the encephalopathy (see Chapter 16 for further details).

Hyponatremia

Hyponatremia has many causes, including liver and cardiac failure, syndrome of inappropriate antidiuresis (neurological causes of which include head trauma, bacterial meningitis, encephalitis, cerebral infarction, subdural and subarachnoid haemorrhage), vomiting and diuretic use. The severity of the symptoms is related to the rapidity of decline in plasma sodium. This is because hyponatremia causes extracellular hyposmolarity and a tendency for free water to shift from the vascular space to the intracellular space. When plasma sodium concentration falls slowly, over a period of several days or weeks, the brain is capable of compensating by extrusion of solutes to the extracellular space. This reduces the flow of free water into the intracellular space and symptoms are much milder for a given degree of hyponatraemia. When the plasma sodium concentration falls rapidly, this compensatory mechanism is overwhelmed and severe cerebral oedema may ensue. For example, a sodium concentration falling from normal to < 110 mmol/L over 24–48 h will often cause coma. It is necessary to correct the plasma sodium concentration in a controlled manner because of the risk of central pontine myelinolysis if the increase is too rapid (see Chapter 4).

Central pontine myelinolysis is characterized by quadriparesis, dysphagia, dysarthria, diplopia and altered consciousness. Areas of the brain other than the pons are also often involved. Magnetic resonance imaging reveals signal change within these areas but the appearances may be normal early in the course of the disease. Central pontine myelinolysis may prove fatal, and incomplete recovery is common in survivors.

Hypernatraemia

Common causes of hypernatraemia include diabetes insipidus, hyperglycaemia, diarrhoea and fluid deprivation. Again, it is the rate of change of sodium concentration that is important rather than the actual values. A sodium concentration of > 160 mmol/L can cause coma due to cerebral cellular dehydration if it develops rapidly. Treatment depends on the cause of the disorder and the total body fluid volume (see Chapter 4).

Hypercalcaemia

Severe hypercalcaemia (> 3.4 mmol/L) can lead to coma, preceded by fatigue, headache, muscle weakness, irritability and confusion. High concentrations of calcium ions decrease neuronal excitability, which together with dehydration is thought to cause the neurological problems (see Chapter 6).

Septic encephalopathy

Severe extracranial infection can be associated with an impaired mental state. The pathophysiological mechanism involves reduced cerebral blood flow and oxygen extraction by the brain, cerebral oedema, disruption of the blood–brain barrier owing to inflammatory mediators acting on the cerebrovascular endothelium, and

abnormal neurotransmitter compositions. Diagnosis of sepsis can be aided by the finding of a raised white blood cell count and raised inflammatory markers, and by blood cultures. Treatment is of the underlying infection with supportive care.

Autoimmune encephalopathy

The last decade has seen significant advances in our understanding of antibody-mediated encephalopathies. There are two types of antibodies detectable in blood in these conditions (Table 36.5): paraneoplastic antibodies to intracellular targets that are associated with cancer but are not pathogenetic, and antibodies to the extracellular domain of neuronal cell-surface proteins that directly cause encephalitis and are also often associated with cancer. Autoimmune encephalitis typically presents with subacute memory loss, psychiatric and behavioural disturbance and seizures. The detection of one of the specific antibodies supports a diagnosis of paraneoplastic encephalitis. Treatment is of the underlying cancer and with immunotherapy.

Hashimoto encephalopathy presents with seizures, behavioural and psychiatric manifestations, movement disorders and coma and is associated with the presence of high titres of anti-thyroglobulin or anti-thyroid peroxidase (TPO, antimicrosomal) antibodies. The condition can present without any features of thyroid disease. It is unclear whether the anti-thyroid antibodies are pathogenetic or simply represent an immune epiphenomenon but it is important not to miss this diagnosis as the condition responds to treatment with steroids.

Dementia

Dementia is a syndrome characterized by progressive deterioration of cognitive function without alteration of consciousness. Cognitive deficits most commonly affect memory, but other cognitive domains such as language,

praxis, visual perception and most notably executive function are also often affected. Diagnosis of dementia is generally by clinical criteria during life, although many of its causes are defined on the basis of histopathological criteria with the consequence that definitive diagnosis can only be made at post-mortem. However, over the last few years a number of laboratory CSF biomarkers have emerged and have an increasing role in diagnosis. Cerebrospinal fluid A β -40, A β -42, total tau and phosphorylated tau proteins are the most sensitive biomarkers for the diagnosis of Alzheimer disease and 14-3-3 protein for the diagnosis of Creutzfeldt-Jakob disease (see Chapter 34).

SPINAL CORD DISORDERS

Spinal cord disease usually presents with motor and/or sensory symptoms in either the upper and lower or just the lower limbs, depending on the site of the insult. Clinical signs include increased tone in the limbs, brisk reflexes and, typically, well-preserved muscle bulk. Bowel and bladder function can be compromised. The list of possible causes of spinal cord disease is extensive and includes vascular, degenerative, compressive, inflammatory, malignant, metabolic and infectious conditions. Metabolic causes include vitamin B₁₂ deficiency (Box 36.2) and adrenomyeloneuropathy. A normal MRI scan of the spinal axis will exclude a significant number of differential diagnoses.

Vitamin B₁₂ deficiency (subacute combined degeneration of the spinal cord)

Deficiency of vitamin B₁₂ can produce haematological and neurological abnormalities: features of gastrointestinal disease may also be apparent (see Chapter 12). From an early stage, it is the spinal cord that is predominantly affected. The onset of symptoms is insidious, with symmetrical, uncomfortable, tingling paraesthesiae, initially in the feet but later involving the hands. As the condition progresses, the gait becomes ataxic (sensory ataxia owing to loss of proprioception) and the limbs become weaker and increasingly spastic. Additional problems include peripheral polyneuropathy, optic neuropathy, psychiatric disturbance and dementia.

The exact pathophysiological process leading to neuronal damage in vitamin B₁₂ deficiency is unknown. Vitamin B₁₂ is involved as a cofactor in the conversion

TABLE 36.5 Antibody-mediated encephalopathy

Antibody	Associated tumour
Paraneoplastic antineuronal antibodies (intracellular)	
Anti-Hu	Bronchial small cell carcinoma (SCLC)
Anti-Ma2 (anti-Ta)	Testicular tumour
Anti-collapsin response mediator protein 5 (anti-CV2)	Lymphoma, SCLC
Antineuronal nuclear 3	SCLC
Autoimmune receptor antibodies (extracellular)	
Voltage-gated potassium channel	Uncommon
AMPA receptor (glutamate receptor 1,2 subunits)	SCLC, breast, thymoma (70%)
γ -Aminobutyric acid B-receptor	SCLC (47%)
Glutamic acid decarboxylase receptor	Rare
N-methyl-D-aspartate receptor	Ovarian teratoma (20–59%), rarely testicular teratoma or SCLC.

AMPA, amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid.

BOX 36.2 Causes of metabolic spinal cord disorders

- Vitamin B₁₂ deficiency
- Folate deficiency
- Copper deficiency (copper deficient myelopathy)
- Vitamin E deficiency
- Toxic myelopathy (e.g. lathyrism, fluorosis, drugs)
- Hepatic myelopathy
- Metabolic myelopathy of genetic origin
 - Hexosaminidase A deficiency
 - Adrenomyeloneuropathy

of homocysteine to methionine and of methylmalonyl-CoA to succinyl-CoA. It is possible that the impaired synthesis of methionine leads to a depletion of S-adenosylmethionine, which is required for myelin synthesis.

The diagnosis is confirmed by measuring vitamin B₁₂ concentration in the serum. In some patients, the concentration may be only 'low-normal'. In borderline cases, measurement of homocysteine and methylmalonate concentrations can be useful. It is important that the diagnosis is made as early as possible as this is a potentially reversible condition. Magnetic resonance imaging of the spine often shows the spinal cord to be atrophic and can show signal change in the dorsal columns.

Folate deficiency

Folate deficiency is usually associated with other nutritional deficiencies. Folate antagonists (e.g. methotrexate, trimethoprim and pyrimethamine) are also potential causes. Neurological manifestations are rare and the same as with vitamin B₁₂ deficiency. Treatment is by folate replacement, although any vitamin B₁₂ deficiency should be treated first to prevent exacerbation of spinal cord degeneration.

Copper deficiency

Copper deficiency is rare and causes both haematological and neurological disease. Neurological manifestations include myelopathy, peripheral neuropathy and optic neuropathy. Copper deficiency was recognized in ruminants (swayback disease) long before it was identified in humans. The presentation can be similar to that of subacute combined degeneration of the spinal cord. It is usually secondary to gastric surgery, gastrointestinal diseases or total parenteral nutrition, or to excess intake of zinc or iron, which reduce copper absorption in the gut. Copper deficiency myelopathy has been attributed to ingestion of zinc contained in denture cream in some patients. Treatment is with oral or intravenous copper replacement.

Vitamin E deficiency

Vitamin E deficiency is rare, and causes a spinocerebellar ataxia, retinopathy, myopathy and anaemia. Poor diet is not usually sufficient cause unless associated with malabsorption. Abetalipoproteinemia is a rare inherited disorder of fat metabolism that results in poor absorption of dietary fat and vitamin E. Vitamin E deficiency is discussed further below (see ataxia).

Hepatic myelopathy

Hepatic myelopathy is a rare complication of chronic liver disease. It has been suggested that it is caused by the neurotoxicity of ammonia or other metabolites bypassing normal liver metabolism. It is characterized by spastic paraparesis with minimal sensory and sphincteric involvement. Diagnosis is a process of exclusion requiring normal CSF analysis and brain imaging. Liver

transplantation may result in improvement, especially if performed early in the course of the disease.

Hexosaminidase A deficiency

This is a progressive neurodegenerative disease with variable age of onset and rate of progression and hence variable prognosis. The clinical presentation of the chronic form of hexosaminidase A deficiency may mimic spinocerebellar degeneration, Friedreich ataxia, or amyotrophic lateral sclerosis (see below).

Adrenomyeloneuropathy

This is the commonest clinical variant of adrenoleukodystrophy, a disorder of peroxisomal fatty acid oxidation causing the accumulation of very long chain fatty acids in myelin, adrenal cortex and Leydig cells of the testes. Patients with an adrenomyeloneuropathy presentation of this X-linked recessive disorder present in their third decade with a progressive spastic paraparesis and sphincter dysfunction. A sensorimotor peripheral neuropathy may be a feature and ataxia and dementia are seen occasionally. Adrenal insufficiency will often have been present since childhood and there may be hypogonadism. Around 50% of heterozygous females will show some symptoms later in life. The clinical findings and magnetic resonance imaging point to the diagnosis, which is confirmed by demonstrating hypoadrenalism and elevated concentrations of very long chain fatty acids (VLCFAs) in plasma and cultured skin fibroblasts. Genetic testing is available. The only gene in which mutations are known to cause adrenoleukodystrophy is *ABCD1*, which codes for a peroxisomal membrane transporter protein, although molecular genetic testing is rarely required to confirm the disease, especially in a male. Treatment is with steroid replacement, symptomatic management and supportive care.

PERIPHERAL NEUROPATHY

Peripheral nerve disorders are the most common neurological diseases and their prevalence has been estimated to be 2–8% of the adult population. There are many causes of peripheral neuropathy and the range of clinical presentation is wide. Peripheral neuropathy can be categorized by the clinical pattern: mononeuropathy (e.g. carpal tunnel syndrome); multiple mononeuropathy (e.g. mononeuritis multiplex in vasculitis), or distal symmetrical polyneuropathy (e.g. neuropathy in diabetes mellitus). With the help of electrodiagnostic testing (nerve conduction studies and electromyography), neuropathies can be further categorized as demyelinating, axonal or mixed type of neuropathy. Appropriate laboratory tests may help to identify a specific diagnosis. Neuropathies with a specific metabolic basis are listed in [Box 36.3](#). Peripheral neuropathies of toxic or metabolic aetiologies are usually symmetrical, distal, sensorimotor and axonal. The American Academy of Neurology has published practice parameters to guide laboratory and genetic testing in distal symmetric polyneuropathy (see [Further reading](#)).

BOX 36.3 Metabolic causes of peripheral neuropathy

- Diabetes
- Chronic kidney disease
- Liver disease
- Endocrine
- Immune mediated and paraproteinaemic
- Nutritional deficiency
- Vitamin deficiency
 - Vitamin B₁₂
 - Thiamin
 - Pyridoxine
 - Vitamin E
 - Niacin (vitamin B₃)
 - Pantothenic acid (vitamin B₅)
 - Folic acid
- Neuropathy associated with bariatric surgery
 - Hypophosphataemia
 - Copper deficiency
 - Strachan syndrome
- Metabolic
 - Refsum disease
 - Porphyria
 - Fabry disease
 - Cerebrotendinous xanthomatosis
 - Tangier disease
 - Amyloid
 - Mitochondrial disorders
- Paraneoplastic

Neuropathy can present with a variety of signs and symptoms. Symptoms can be classified as either positive or negative. Positive symptoms reflect inappropriate nerve activities, whereas negative symptoms reflect reduced nerve activity. Positive sensory symptoms are burning or lancinating pain and paraesthesiae, whilst negative sensory symptoms are numbness, lost of proprioception and difficulty differentiating cold and hot sensations. Weakness and fatigue can be considered as negative motor symptoms, while cramps, muscle twitching and myokymia are positive motor symptoms. The findings on clinical examination of the sensory systems include sensory loss, trophic skin changes, sensory ataxia and Charcot joints. Features of motor dysfunction are dominated by lower motor neuron wasting, normal tone, absent reflexes and flexor plantar responses. Peripheral nerves should also be palpated to check for thickening, which occurs in leprosy, amyloidosis and acromegaly. Gastrointestinal and cardiovascular symptoms can be prominent features in autonomic dysfunction, causing bloating, constipation, diarrhoea, light headedness, palpitations and excessive sweating. Blood pressure should be checked with the patient lying and standing to document a postural drop caused by autonomic dysfunction. Neuropathies with a metabolic cause tend to be symmetrical, other than for some types related to diabetes, and with variable onset.

The investigations should be guided by the clinical picture. Polyneuropathy can arise in the course of many illnesses, particularly diabetes (Table 36.6). Initial investigations would reasonably include the tests listed in Box 36.4 and should yield a diagnosis in 60–90% of patients. The majority of patients will have diabetes or alcohol exposure.

TABLE 36.6 Most common causes of a symmetrical neuropathy

Condition	Proportion of symmetrical neuropathies (%)
Diabetes	11–41
Chronic idiopathic axonal neuropathy	10–40
Paraproteinaemia	9–10
Alcohol misuse	7
Chronic kidney disease	4

BOX 36.4 Initial investigation of patients with peripheral neuropathy

- Full blood count and erythrocyte sedimentation rate (ESR)
- Fasting glucose
- Thyroid function tests
- Creatinine
- Liver function tests
- Vitamin B₁₂ with metabolites (methylmalonic acid with or without homocysteine) if low-normal
- Antinuclear antibodies (ANA)
- Serum protein electrophoresis with immunofixation if required
- Chest X-ray

There are no diagnostically useful antibody assays. More invasive tests (Box 36.5) are best undertaken after consultation with a physician or neurologist with an interest in peripheral neuropathy. The nerve conduction tests will reveal whether the neuropathy is axonal or demyelinating.

Despite all efforts, about one-third of patients with peripheral neuropathies remain without a diagnosis. About 40% of these will be found to have impaired glucose tolerance, although how much this contributes directly to the development of neuropathy is still debated. Treatment with diet and exercise reduces the rate of progression to diabetes and may also improve the neuropathy, at least in the short term.

BOX 36.5 Further investigation of patients with peripheral neuropathy

- Glucose tolerance test
- Nerve conduction studies, electromyography
- Bence–Jones protein in urine
- Immunological tests for antineuronal antibodies, anti-ganglioside antibodies, markers for vasculitis, anti myelin-associated glycoprotein antibodies, other connective tissue disorders
- Microbiology tests including for Lyme disease; HIV serology in at risk patients
- Cerebrospinal fluid for cells, protein, oligoclonal bands
- Tests for Sjögren syndrome, e.g. salivary flow rate, Schirmer test, Rose Bengal test
- Search for carcinoma, lymphoma or myeloma
- Molecular genetic tests
- Toxicology screening

Patients with non-painful chronic idiopathic neuropathies have been found to have higher blood triglyceride concentrations and more exposure to environmental toxins compared with controls. Currently, it is unclear whether these observations will translate into any therapeutic benefit.

Small fibre painful axonal neuropathy

There is a small subset of patients in whom small nerve fibres alone are affected and in whom pain and temperature sensation but not large fibre modalities (vibration and proprioception) are impaired. Conventional clinical neurophysiological tests are normal because they do not assess small nerve fibre function. Most patients with a small fibre neuropathy have diabetes mellitus or at least impaired glucose tolerance but a few have Sjögren syndrome and other rare causes.

Diabetic neuropathies

Peripheral neuropathy is rarely encountered in young patients with newly diagnosed diabetes mellitus but diabetes presenting in older patients is frequently associated with a neuropathy and this may be the presenting problem. Studies have suggested that after 25 years of diabetes, about 50% of patients will have evidence of a neuropathy, most commonly chronic diabetic peripheral neuropathy. Poor glycaemic control, duration of diabetes, hyperlipidaemia (particularly hypertriglyceridaemia), increased albumin excretion and obesity are risk factors for the development of neuropathy.

Although distal symmetrical peripheral neuropathy is the commonest presentation, the manifestations of diabetic neuropathy are varied and can be subdivided clinically into the symmetric polyneuropathies and the focal or multifocal neuropathies (Box 36.6).

Symmetrical polyneuropathies

The onset of distal sensory and sensorimotor polyneuropathy is insidious, with mild symptoms of numbness and paraesthesiae affecting the toes and feet. A burning pain, worse at night, may be a prominent symptom. The typical findings on examination are sensory loss to pinprick, light touch and vibration sensation in a stocking distribution, with absent ankle reflexes. Occasionally, there may be mild distal weakness. Foot ulceration and neuropathic arthropathy are potentially serious complications related to loss of pain sensation. Ulcers tend to be

situated at pressure points – over the toes, heels, malleoli and metatarsal heads. In some patients, neuropathy may be precipitated by the institution of therapy with either insulin or oral hypoglycaemics and even dietary control. The neuropathy then becomes evident 4–6 weeks after starting treatment, with painful paraesthesiae affecting the lower half of the legs.

Acute painful neuropathy is a distinct entity characterized by marked weight loss over a short period of time followed by severe burning pains in the feet with contact discomfort. By maintaining good glycaemic control, patients slowly improve.

Proximal symmetrical motor neuropathy and asymmetrical proximal neuropathy (see below) are now termed lumbosacral radiculoplexus neuropathy (diabetic amyotrophy). There is often diffuse lower back pain at the onset, followed by progressive proximal weakness and wasting of the muscles. The knee reflexes are absent but sensory loss is minimal. Good glycaemic control leads to slow recovery over 12–30 months.

Focal and multifocal neuropathies

An isolated third nerve lesion is the commonest of the cranial nerve palsies, usually with sparing of pupillary function; in about half of patients, there is aching retro-orbital pain. The lesion is thought to be an infarct in the nerve trunk (i.e. it has microvascular aetiology).

Diabetic patients are more prone to isolated peripheral nerve palsies (mononeuropathy), suggesting increased susceptibility to compressive injury as the nerves are damaged at sites of pressure (e.g. the ulnar nerve at the elbow, the median nerve at the wrist).

Lower limb asymmetrical proximal motor neuropathy, also called lumbosacral radiculoplexus neuropathy, is probably caused by an underlying microvasculitis. The onset is usually subacute, with thigh pain, followed by asymmetric weakness and wasting of the lower limb muscles, most often affecting hip and knee flexors. Sensory loss is minimal. The weakness can be very severe. Recovery is slow, occurring over 2–4 years and is often incomplete.

Pathophysiology of diabetic neuropathy

The pathophysiology of diabetic neuropathy is only partially understood. A combination of direct axonal injury from hyperglycaemia, insulin resistance, central adiposity (visceral fat produces toxic adipokines), endothelial injury and microvascular dysfunction is possibly responsible. Hyperglycaemia is known to increase flux through the polyol pathway, increase nonenzymatic glycation of proteins and enhance oxidative stress.

Polyol pathway. Aldose reductase converts glucose to sorbitol, which is then converted to fructose by sorbitol dehydrogenase. In nerves from diabetic patients, accumulation of sorbitol leads to a compensatory reduction in other osmolytes such as myoinositol and taurine. The loss of myoinositol may lead to a reduction in nerve Na^+, K^+ -ATPase, which then causes nerve conduction defects. Aldose reductase inhibitors showed early promise by reversing the biochemical changes in experimental

BOX 36.6 Presentations of diabetic neuropathy

- Symmetrical polyneuropathy sensory/sensorimotor and autonomic neuropathy
- Lumbo-sacral plexopathy
- Focal and multifocal neuropathies
- Cranial nerve palsies
- Trunk and limb mononeuropathy
- Asymmetric lower limb motor neuropathy
- Thoracoabdominal neuropathy

diabetic nerves but, in clinical trials, these compounds have had no significant impact on human diabetic neuropathy.

Non-enzymatic glycation. The formation of irreversible advanced glycation end products of various components in the nerves of patients with diabetes is believed to contribute to the structural changes that are seen in diabetic neuropathy. Myelin components and axonal cytoskeletal proteins are glycated, leading to impaired cytoskeletal assembly with consequently abnormal axonal transport and axonal atrophy. Aminoguanidine, developed as a potential inhibitor of glycation, causes improvement in Na^+, K^+ -ATPase activity and nerve conduction velocity when administered to experimental diabetic rats. It has not been tested in humans.

Oxidative stress. Diabetes is associated with the increased production of free radicals, which have the potential to cause damage to arteries. Endothelial cell function is altered and this may impair nerve blood flow and nerve oxygenation. Thioctic acid (α -lipoic acid) reduces oxidative stress, and is available in some countries for the treatment of diabetic patients with neuropathic pain.

Immune mediated neuropathies

Immune mediated neuropathies are heterogeneous disorders, with a wide range of presentations. The pathological process affects the nervous system either directly (e.g. demyelination) or indirectly (e.g. secondary to vasculitis).

Acute inflammatory neuropathies and variants

Guillain–Barré syndrome is an acute immune mediated polyradiculoneuropathy, usually with demyelination. In two thirds of patients it is preceded by an infection. The presentation is acute and evolves over a few weeks, typically with an ascending flaccid paralysis, early loss of deep tendon reflexes and minimal sensory signs. Sphincteric functions are usually spared. Significant autonomic features are common and neuropathic pain may be problematic. The salient feature of CSF analysis is elevated protein with a normal cell count. The finding of an increased cell count warrants further investigation for conditions such as human immunodeficiency virus (HIV) and Lyme disease. Antibodies associated with Guillain–Barré syndrome are shown in Table 36.7.

A variety of infections can precede the onset of Guillain–Barré syndrome and these can be associated

with various anti-ganglioside antibodies. *Campylobacter jejuni* is associated with GM1, GM1b, GD1a, GalNAc-GD1a, GD3, GT1a and GQ1b. *Haemophilus influenzae* is associated with GM1 and GT1a. *Mycoplasma pneumoniae* and cytomegalovirus are associated with galactocerebroside GM2 antibody.

The mainstay of treatment for Guillain–Barré syndrome is intravenous immunoglobulin or plasmapheresis, usually in an intensive care setting.

Chronic inflammatory demyelinating polyneuropathies and variants including paraproteinaemic neuropathies

Chronic inflammatory demyelinating polyneuropathy (CIDP) is a heterogeneous group of acquired immune-mediated neuropathies. It may present in isolation or as part of a systemic disease. The classic form of CIDP is symmetrical with motor involvement greater than sensory. Weakness is present in both proximal and distal muscles. It can be associated with various systemic diseases such as HIV infection, connective tissue disorders and paraproteinaemia (see Chapter 30). Serum protein electrophoresis is therefore one of the investigations that should be carried out routinely in all adult patients presenting with a peripheral neuropathy. Treatment of CIDP includes intravenous immunoglobulin, plasmapheresis and immunomodulating treatment including steroids.

Monoclonal gammopathy of unknown significance

This can cause a neuropathy that is slowly progressive over many years and has a symmetrical distribution, first affecting the feet with numbness and paraesthesiae but later involving the hands. Muscle weakness and wasting occurs months or years after the onset of sensory symptoms. The cause of the neuropathy is unknown. The commonest paraprotein to be detected is IgM, which can bind to peripheral nerves via the carbohydrate moiety in myelin-associated glycoprotein (MAG). Antibodies to MAG are found in 60% of patients. IgG and IgA neuropathies are less common, but are more responsive than IgM neuropathy to immunotherapy and plasmapheresis.

Multiple myeloma

This rarely causes an isolated neuropathy. When it does occur, it is sensorimotor in type and may be axonal and demyelinating. Treatment of the myeloma may stabilize the neuropathy.

TABLE 36.7 Antibodies associated with Guillain–Barré syndrome

Clinical syndrome	Anti-ganglioside antibody	Isotypes	Comment
Acute inflammatory demyelinating polyneuropathy	LM1	IgG	Present in 5–20% of patients. Not clinically useful
Acute motor (and sensory) axonal neuropathy	GM1, GM1b, GD1a	IgG	
Miller–Fisher variant of Guillain–Barré syndrome	GQ1b, GT1a	IgG	Highly specific and sensitive, found in 90% of patients

Adapted from Willison HJ, Yuki N. Peripheral neuropathies and anti-glycolipid antibodies. Brain 2002; 125: 2591–2625.

Waldenström macroglobulinaemia

This is commonly associated with a sensorimotor neuropathy in addition to the more generalized symptoms of fatigue, weakness and a bleeding diathesis.

POEMS syndrome

Polyneuropathy, organomegaly, endocrinopathy, monoclonal gammopathy and skin changes (POEMS syndrome) has a peak incidence at age 50–60 years. It is distinct from the neuropathy associated with myeloma and its pathogenesis remains unknown. The main clinical features are of a progressive sensorimotor neuropathy associated with a monoclonal plasma cell proliferative disorder and osteosclerotic bone lesions. The light chain is invariably λ rather than κ , and when present, the heavy chain is either A or G and only occasionally M. The endocrine disturbance can appear at any stage of the illness and can involve adrenal, thyroid, pituitary, parathyroid, gonadal or pancreatic function. There is no specific diagnostic test for POEMS syndrome.

Chronic kidney disease and established renal failure

Peripheral neuropathies are not uncommon in patients with chronic kidney disease (CKD). A variety of peripheral nerve disorders may occur: carpal tunnel syndrome, ulnar and femoral neuropathy and uraemic polyneuropathy. The pathogenesis probably involves anatomical vulnerability, ischaemia, deposition of β_2 -microglobulin-associated amyloid and tissue compression around nerves.

The neuropathy associated with non-diabetic renal disease is of insidious onset, progressing over months and is present in 60–100% of patients on dialysis. Neuropathy generally develops only at CKD stage 5 and is characterized by paraesthesiae, tingling and burning of the feet, cramps and restlessness of the legs. Weakness is uncommon and, when present, is relatively mild. Patients may also develop autonomic features, with postural hypotension, impaired sweating, diarrhoea, constipation or impotence. Although there is a correlation between the impairment of nerve conduction and decreased glomerular filtration rate, starting dialysis or increasing its rate does not seem to improve an established neuropathy. However, improvement is seen after renal transplantation. The neuropathy has been attributed to the inability of dialysis membranes to clear ‘middle molecules’ (see Chapter 7), which are thought to be neurotoxic. The evidence to support this hypothesis is conflicting.

Liver disease

The prevalence of peripheral neuropathy in chronic liver disease is variable. Alcohol is the commonest cause of neuropathy associated with liver disease owing both to the direct neurotoxic effect of alcohol and to liver failure. Peripheral neuropathy can occur as a direct result of a primary liver disease although the underlying mechanism of neuropathy is unclear. It can also be part of a systemic disease that affects both the liver and peripheral

nervous system for example, vasculitis. Most patients with chronic liver disease have evidence of an axonal neuropathy. There may be an additional autonomic neuropathy, which carries a poor prognosis, and carpal tunnel syndrome is common. Acute peripheral neuropathy, especially Guillain–Barré syndrome, can be associated with viral hepatitis. Primary biliary cirrhosis is associated with a painful sensory neuropathy with sensory ataxia.

Endocrine disturbances

Hypothyroidism

Hypothyroidism frequently causes carpal tunnel syndrome because of deposition of mucopolysaccharide complexes within the carpal tunnel. Adequate thyroxine replacement is often the only treatment necessary.

Hyperthyroidism

Hyperthyroidism can also cause carpal tunnel syndrome that typically resolves with treatment. The mechanism is thought to be the same as in hypothyroidism. Thyroid storm can be associated with acute neuropathy.

Acromegaly

Acromegaly causes swelling of the median nerve from oedema and/or hypertrophy, resulting in an increased incidence of carpal tunnel syndrome. Treatment of the acromegaly improves the symptoms. There can also be a polyneuropathy, which is often painful. The relationship between peripheral neuropathy and excess growth hormone is not clear and can be complicated by coexisting diabetes mellitus.

Nutritional peripheral neuropathies

Nutritional deficiencies and excess alcohol are associated with peripheral neuropathies as well as encephalopathy and spinal cord disorders (see previous sections).

Vitamin B₁₂ deficiency

This is commonly associated with peripheral neuropathy, which may be precipitated by nitrous oxide anaesthesia. Measurement of plasma vitamin B₁₂ concentration may give a low-normal value even though there is a true functional deficiency. It is therefore recommended that, for patients with a neuropathy and borderline B₁₂ value, plasma methylmalonate and homocysteine concentrations should also be checked. In one review, patients with abnormal methylmalonate and homocysteine concentrations accounted for 15% of those previously labelled as ‘cryptogenic neuropathy’ as the plasma vitamin B₁₂ concentrations were within reference limits. Interestingly, a megaloblastic anaemia is rarely seen in patients with a vitamin B₁₂ deficiency state and neurological complications.

Intramuscular injection of vitamin B₁₂ is the standard treatment for such deficiency, but the recovery of neuropathy is variable, and may not occur at all.

Thiamin (vitamin B₁) deficiency

Also known as beri-beri, this is now rarely encountered in developed countries; when it does occur, it is most often associated with alcoholism and severe anorexia. Thiamin and its coenzyme thiamin pyrophosphate are required for mitochondrial function and myelin formation. A deficiency state has to be more prolonged before the peripheral nerves become affected than is the case with encephalopathy. The symptoms are numbness and pain that start in the feet and may later involve the hands. Although direct measurements of thiamin concentration are available, their use in diagnosing thiamin deficiency is controversial. If deficiency is suspected clinically, treatment should be given immediately either orally or intravenously. Reversal of the neuropathy is unpredictable and it may take 12 months for any improvement to be seen.

Vitamin B₆ (pyridoxine) deficiency

This deficiency causes an axonal neuropathy that develops very slowly and typically results in distal numbness and mild weakness. In contrast, a severe ataxic painful neuropathy can result from vitamin B₆ toxicity at doses of more than 200 mg/day.

Vitamin E deficiency

Vitamin E deficiency only rarely causes an isolated neuropathy; measuring plasma vitamin E concentrations should not be part of the routine screen in neuropathy. Chronic vitamin E deficiency in developed countries is usually associated with disorders of lipid absorption. Vitamin E has a role in maintaining cell membrane structure and as a free radical scavenger. After long exposure to low concentrations of vitamin E, ataxia and axonal neuropathy may develop. The neurological features can improve with oral vitamin E treatment and it is also recommended that high doses of vitamin A should be given at the same time.

Niacin (vitamin B₃), pantothenic acid (vitamin B₅) and folic acid deficiencies

These are also associated with peripheral neuropathy.

Chronic hypophosphataemia

This is seen most commonly in patients on long-term parenteral nutrition and may result in a Guillain-Barré like syndrome.

Copper deficiency

Copper deficiency may be associated with gastric surgery and may mimic subacute combined degeneration of the cord (see p. 689).

Neuropathy associated with bariatric surgery

Untreated, up to 62% of patients who have had bariatric surgery develop a peripheral neuropathy. The neuropathy is predominantly sensory. Some 31% develop features

of encephalopathy, and 28% have Wernicke-Korsakoff syndrome. The onset varies from 3 to 20 months following surgery; outcomes are mixed. The most frequent associated deficiencies are iron, thiamin, folate, vitamin D and calcium. Delayed (>2 years after surgery) deficiency of copper has also been reported. Current guidelines recommend trace element and vitamin replacement and careful monitoring after surgery to minimize the development of neurological and other complications.

Strachan syndrome

Previously termed Jamaican neuritis, Strachan syndrome was first identified in the West Indies during periods of famine. It is characterized by a painful, predominantly sensory, ataxic neuropathy, and may also be associated with optic neuropathy, sensorineural hearing loss, dorsal lateral myelopathy, glossitis and stomatitis. It is thought to be caused by multiple deficiencies of thiamin, niacin, riboflavin, pyridoxine and cobalamin, and hence treatment is administration of multivitamins, although recovery may only be partial.

Metabolic neuropathies**Refsum disease (hereditary ataxia polyneuritis)**

This is an autosomal recessive metabolic disease, characterized by accumulation of phytanic acid in plasma and tissues. It is caused by deficiency of peroxisomal phytanoyl-CoA hydroxylase (phytanic acid oxidase), which converts phytanic acid to α -hydroxyphytanic acid; mutations in the gene for this enzyme (located on chromosome 10p13) have been identified in most patients, although other genes may also be involved. A progressive symmetrical demyelinating sensorimotor neuropathy, cerebellar ataxia, retinitis pigmentosa initially presenting with night blindness, and ichthyosis are hallmarks of the disease. The age of onset of the disease can vary from childhood to adulthood, and as the condition progresses, the patient develops an ataxic gait, distal limb weakness and numbness. Other important clinical features include a 'salt and pepper' type of retinal pigmentation (flecks of dark pigment with areas of whitish depigmentation), nerve deafness, cataracts and cardiomyopathy. Skeletal deformities, including kyphoscoliosis, pes cavus, and shortened metacarpals and metatarsals, are also encountered. Optic atrophy, cataract and vitreous abnormalities cause further deterioration of vision.

The neuropathy starts distally with marked wasting and weakness. Symptoms of pain and paraesthesiae occur in some but not all patients. The tendon reflexes are lost as the condition progresses and there is loss of sensation (especially vibration sensation and joint position sense) in a glove and stocking distribution. The peripheral nerves may be enlarged. Without treatment, there is gradual deterioration, but about half of those affected may enter a phase of remission that can last for several years.

The diagnosis of Refsum disease is made on the basis of clinical findings and a plasma phytanic acid concentration >200 μ mol/L. However, plasma accumulation of phytanic acid is not specific for the condition as raised

concentrations may occur in other peroxisomal disorders. In addition, patients have been recorded with defective α -oxidation but normal phytanic acid concentrations. Therefore, to be certain of the diagnosis of Refsum disease, it has to be shown that there is a defect in α -oxidation. This can be done by measuring the production of $^{14}\text{CO}_2$ from a cultured skin fibroblast monolayer using (1- ^{14}C) phytanic acid as substrate. Cerebrospinal fluid protein concentration is often raised. Molecular genetic testing is available.

Phytanic acid is derived exclusively from the diet, therefore dietary control is the treatment of choice. This requires the elimination, as far as possible, of all dairy products and any fatty meat or oily fish. However, such a diet is protein and energy poor so liquid nutritional supplements are often required. In those who respond with a fall in plasma phytanic acid concentrations, progression of the neurological damage is halted, and in those with neuropathy, there is improvement in nerve conduction velocities. As dietary manipulation takes time to correct plasma phytanic acid concentrations, weekly plasma exchange can be used to attain normal concentrations quickly for patients with more rapid clinical progression.

Porphyric neuropathy

Porphyria is a rare but important cause of peripheral neuropathy. Its diagnosis depends on demonstrating an abnormality in haem metabolism (see Chapter 28). Only the hepatic porphyrias (acute intermittent porphyria, AIP, hereditary coproporphyria and variegate porphyria) are associated with neuropathy, the most severe attacks being associated with AIP. Within 2–3 days of the onset of an acute attack, the neuropathy begins with back and limb pains and the early development of marked symmetrical limb weakness, which may spread to involve the cranial nerves and respiratory muscles. The patient may have paraesthesiae and the sensory loss may be either in a symmetrical glove and stocking distribution or patchy. Autonomic features can also be present – hypertension, tachycardia, constipation and bladder disturbance. Other clinical problems such as seizures or psychiatric disturbance and a normal CSF protein may help to distinguish porphyric neuropathy from the Guillain-Barré syndrome. The exact pathogenetic mechanism is not clear but evidence suggests that both direct toxic effects of 5-aminolaevulinic acid and intracellular metabolic derangements, including reduced activity of Na^+, K^+ -ATPase, contribute to the neurological disorders.

Treatment of an acute episode includes avoiding any drugs that may exacerbate the condition, and appropriate sedation and pain relief. A glucose infusion should be commenced. The use of specific measures, for example haematin, is discussed in Chapter 28. Recovery is usually good but may be incomplete in some patients.

Fabry disease (*angiokeratoma corporis diffusum*; α -galactosidase deficiency)

This is an inherited X-linked recessive disorder of glycosphingolipid metabolism resulting from a deficiency of the enzyme ceramide trihexosidase (α -galactosidase).

The gene for this enzyme (GLA) lies on the long arm of the X chromosome between Xq21.33 and Xq22. However, females heterozygous for the condition have low enzyme activity and can be mildly affected clinically. The accumulating ceramide trihexoside consists of sphingosine (a long chain amino alcohol) linked to a long chain fatty acid. The trihexose (glucose–galactose–galactose) is linked to the primary alcohol group on C1 of ceramide. Because of the low ceramide trihexosidase activity, cleavage of the terminal galactose from the ceramide trihexoside does not occur.

The clinical effects are produced by the deposition of ceramide trihexoside (globotriaosylceramide) in the vascular endothelium. The characteristic manifestations are renal impairment, neuropathic pain and cutaneous angiokeratomas in a bathing trunk distribution. The onset of symptoms is in late childhood or early adolescence, with episodes of severe shooting and burning pains and paraesthesiae in the feet and hands (Fabry crisis), which may be severe enough to confine the patient to a wheelchair. Exercise, fatigue and emotional stress are recognized precipitants of acute painful episodes. The symptoms subside as the patient gets older. In contrast, the characteristic purple skin lesions, corneal dystrophy and renal involvement causing hypertension and CKD become more evident with age. Involvement of the myocardium and mitral valve disease (incompetence and prolapse) are very common, but established renal failure is the usual cause of death. Clinical examination reveals only minimal abnormalities suggesting a peripheral neuropathy. Anhidrosis may suggest autonomic involvement and nerve conduction studies show a mild axonal neuropathy with a selective loss of small myelinated fibres and unmyelinated axons.

Diagnosis in males is reliably made by measuring ceramide trihexosidase activity in plasma, leukocytes or cultured skin fibroblasts. Blood spots can be used for screening at-risk populations. Enzyme activity can also be measured in duodenal or jejunal biopsy material. Glycolipid can be measured in urine sediment. Molecular genetic testing is the most reliable method for the diagnosis of carrier females. Predictive testing is available for at-risk asymptomatic adults and for high risk pregnancies. Standard treatment is enzyme replacement therapy. There have been no human trials of gene replacement to date.

Cerebrotendinous xanthomatosis (*cholestanolosis*)

Cerebrotendinous xanthomatosis is a rare, autosomal recessive lipid storage disease with multiorgan involvement. The clinical manifestations usually start at infancy and develop during the first and second decades of life. Peripheral neuropathy, especially the subtype of axonal sensorimotor neuropathy, is common. Patients may also develop a progressive cerebellar ataxia, juvenile cataracts, chronic diarrhoea and tendon xanthomas. A deficiency of hepatic mitochondrial 27-hydroxylase has been identified. Biochemical findings include high plasma and tissue cholestanol concentration, normal to low plasma cholesterol concentration, decreased chenodeoxycholic acid, increased concentration of bile alcohols

and their glycoconjugates and increased concentrations of cholestanol and apolipoprotein B in cerebrospinal fluid. Molecular genetic testing for *CYP27A1*, the only gene in which mutations are known to cause cerebrotendinous xanthomatosis, is available.

Treatment with chenodeoxycholic acid inhibits abnormal bile acid synthesis and is most effective in reducing elevated plasma cholestanol concentrations and reducing bile alcohols. The use of hydroxymethylglutaryl-CoA (HMG-CoA) reductase inhibitors (statins) may offer some help in slowing the neurological progression.

Tangier disease

This is a rare autosomal recessive disorder of lipid transport, characterized by a deficiency of plasma high density lipoprotein (HDL) and tissue storage of cholesteryl esters, especially in the tonsils, giving them a distinct orange–grey colour (see also Chapter 37). The cause of the neuropathy is not clear, but it seems unlikely that low HDL alone is responsible. Neuropathy occurs in about two-thirds of patients and may be the initial symptom in one half of cases, although in others it is evident only on clinical examination. The low plasma concentrations of HDL are thought to be from increased breakdown. There is hypocholesterolaemia with normal or increased concentrations of triglycerides. There is no specific treatment.

Amyloidosis

There are several types of amyloidosis, including primary, secondary, familial and dialysis-related. The associated constitutional symptoms of weight loss and other organ involvement (e.g. heart or kidneys) should provide the clue to consider amyloid as a cause of neuropathy.

Amyloid neuropathy occurs in approximately 15% of patients with primary amyloidosis, and the median duration of symptoms before diagnosis is 29 months. It can be symmetrical or focal, and can have variable degrees of sensory and/or motor involvement. The neuropathy is chronic, debilitating and relentlessly progressive. Autonomic disturbances also occur. Associated biochemical abnormalities are summarized in [Box 36.7](#).

Secondary amyloidosis occurs in conditions such as chronic infection and rheumatoid arthritis and will not be discussed in detail here.

Familial amyloid polyneuropathy (FAP) is usually dominantly inherited and caused by a mutation in the gene coding for transthyretin (TTR, prealbumin), a protein that is produced in the liver and normally functions as a carrier molecule for thyroxine and retinol. As the same mutation is found in Swedish, Portuguese and

Japanese patients, it is likely that the original mutation was in the Viking population that travelled to Europe and the Far East.

The diagnoses of primary and familial amyloidosis can only be made by demonstrating amyloid deposits in appropriately selected tissue. Initial samples are obtained from low-risk sites such as abdominal wall subcutaneous fat, bone marrow or rectum, and only if these are normal should an affected organ be biopsied. If TTR amyloid is detected, the gene should be sequenced to confirm the mutation, which will allow appropriate genetic counselling.

In FAP, younger patients with aggressive disease may be considered for liver transplantation, which seems to have a favourable effect on the course of neuropathy but not on cardiac or ocular lesions. Oral administration of tafamidis meglumine, which prevents misfolding and deposition of mutated TTR, is under evaluation in patients with FAP. High dose melphalan and stem cell transplantation can be effective. Supportive treatment such as management of neuropathic pain and gastroparesis is also very important. Treatment with melphalan and prednisone gives a survival benefit for patients with primary amyloidosis.

Mitochondrial disorders

The term ‘mitochondrial encephalomyopathy’ reflects the major neurological features – central nervous system and muscle – associated with mitochondrial disorders. Muscle disorders and the relevant mitochondrial enzyme defects are discussed in Chapter 33. It is not uncommon to find some minor peripheral nerve involvement in mitochondrial diseases (approximately 30% of patients), but it is rare for it to be the presenting or dominant clinical feature. [Table 36.8](#) summarizes the main associations.

Paraneoplastic neuropathies

The paraneoplastic neurological syndromes occur in <1% of malignancies and can present months or years before a cancer diagnosis. More than 25% of paraneoplastic neurological syndromes present as diseases of the peripheral nervous system, including sensory and autonomic neuropathies and Lambert–Eaton myasthenic syndrome.

BOX 36.7 Biochemical abnormalities associated with primary amyloidosis

- Proteinuria (80%) with elevated plasma creatinine concentration
- Bence–Jones protein (20%)
- Serum paraprotein (60–70%): IgG > IgA > IgM

TABLE 36.8 Neuropathy in mitochondrial disease

Type	Mitochondrial disease
Axonal	MELAS (A3243G mutation). Sensorimotor neuropathy can be acute or subacute
Demyelinating	MNGIE. Symmetrical sensorimotor neuropathy (occasionally axonal change predominates) MELAS

MELAS, mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes; MNGIE, mitochondrial neurogastrointestinal encephalopathy.

TABLE 36.9 Autoantibodies in paraneoplastic syndromes with peripheral neuropathy

Antibody	Malignancy	Neuropathy
Anti-Hu	Small cell carcinoma of bronchus (SCLC)	Sensory Sensorimotor Motor Autonomic
Anti-Ri Anti-Yo	SCLC, breast Breast, ovary	Sensorimotor Sensorimotor Motor
Anti-CV2	SCLC, breast, colon, prostate, thymus, non-Hodgkin lymphoma	Sensorimotor Autonomic
Voltage-gated potassium channel	Thymoma, SCLC, Hodgkin lymphoma	Neuromyotonia

The most common underlying malignancy in paraneoplastic neuropathies is a small cell carcinoma of bronchus, occurring in 50–75% of patients, presenting classically with a subacute sensory neuropathy and associated with anti-Hu antibodies (also known as antineuronal nuclear antigen antibodies or ANNA-1) in the plasma or CSF. In the majority of patients, the neurological involvement extends beyond the peripheral nervous system, and includes brain stem or limbic encephalitis and cerebellar degeneration.

The other nervous system-specific antibodies are outlined in Table 36.9. Although useful to confirm a clinical diagnosis of paraneoplastic syndrome, most of these markers are not specific enough to determine the origin of the cancer. Examination of CSF often shows an elevated protein with oligoclonal bands and an increased lymphocyte count. The tumour remains undetectable in 15% of patients.

The response to immunomodulatory therapies (steroids, immunosuppressants, intravenous immunoglobulin and plasma exchange) is variable.

MOVEMENT DISORDERS

The term ‘movement disorders’ refers to the neurological conditions that have as their predominant feature an excess or a paucity of movement, which may be voluntary or automatic. Disorders in which there is a predominant excess of movement are often referred to as hyperkinesias. Dyskinesia is also often used as a term referring to hyperkinetic involuntary movements. The five major categories of dyskinesias, in alphabetical order, are chorea, dystonia, myoclonus, tics and tremor.

Disorders with a paucity of movement are referred to as the hypokinesias. The characteristic hypokinetic disorder is Parkinson disease, in which there is bradykinesia (a slowness of movement), reduced amplitude of movements, rigidity and loss of postural reflexes. Tremor (a dyskinesia) is also a common feature of Parkinson disease.

Many of the movement disorders are associated with alterations in the basal ganglia or their connections. They

can also arise as a consequence of cerebellar disease, which typically results in impairment of coordination (asynergy, ataxia), problems with judgement of distance (dysmetria) and a kinetic limb tremor.

Myoclonus and many forms of tremors do not appear to be related primarily to basal ganglia pathology, and often arise as a consequence of abnormal neural activity elsewhere in the central nervous system. Myoclonus can also be present in diseases that involve the cerebellum, such as mitochondrial diseases and storage disorders (e.g. sialidosis and ceroid lipofuscinosis).

In the late onset sporadic movement disorders, protein misfolding and aggregation, with neuronal dysfunction and cell death, is the likely underlying pathological process. In diseases with a younger age of onset, genetic abnormalities leading to cell damage (often owing to deficiencies of proteins involved in mitochondrial function) or biochemical abnormalities (such as deficiencies in enzymes involved in dopamine synthesis) as well as metabolic or storage disorders are more likely.

Parkinsonism

Parkinsonism may be caused by many conditions other than Parkinson disease (PD). However, by far the commonest cause of parkinsonism is idiopathic PD, which is the commonest condition seen in adult movement disorder clinics. Whilst a large number of other movement disorders are genetic in aetiology, patients with a monogenic form of PD are rare. Abnormalities in ten genetic loci have so far been identified in PD. Mutations of the *PARK2* gene, which encodes the protein parkin (a component of a ligase complex that has a role in protein degradation) are the commonest autosomal recessive mutations in young onset patients. Mutations of *PARK8*, which result in abnormalities of LRRK2 (leucine-rich repeat kinase 2, a protein kinase) are the commonest dominant mutations identified. There are likely to be a large number of susceptibility genes for the development of PD, the most recently recognized significant one being a glucocerebrosidase gene mutation, which in homozygous individuals causes Gaucher disease.

The clinical syndrome of PD arises predominantly as a consequence of loss of nigrostriatal dopaminergic neurons but other neuronal populations including serotonergic, noradrenergic and cholinergic neuronal transmission are also important and are increasingly recognized as having a likely role in the development of the non-motor features of PD, such as autonomic dysfunction (postural hypotension, abnormal sweating, bladder, bowel and sexual dysfunction), mood disorders, sleep disorders and dementia.

The phenotype of monogenic PD is almost identical to that of sporadic disease; individuals are very responsive to medication but dystonia in the limbs may be present at onset and patients often develop early motor complications. The so-called ‘atypical parkinsonian syndromes’ are much less common than idiopathic PD and are generally not responsive to the usual first line treatment, levodopa, progress more rapidly and lead to a greater loss of independence with a mean survival of 3–9 years.

There are other rarer causes of parkinsonism, including mitochondrial disorders and neurometabolic disorders. Drug-induced parkinsonism from dopamine blocking drugs is an important differential diagnosis in anyone presenting with a relatively acute disorder.

Tremor

Tremor is a rhythmic oscillation of one or more body parts and is probably the commonest type of movement disorder presenting to neurologists. Tremor can be classified according to its phenomenology (rest, action, kinetic, task specific and body part involved), frequency and amplitude or, more specifically, by aetiology (e.g. Parkinson disease, essential tremor or secondary tremor as in thyrotoxicosis). The parkinsonian syndromes and other degenerative disorders are the commonest causes of rest tremor, along with secondary causes, such as toxins and drug therapies. The commonest causes of action or kinetic tremors are enhanced physiological tremor (stress, anxiety and exercise), essential tremor and cerebellar disease.

Dystonia

Dystonic contractions of muscles have a longer duration than myoclonus or chorea, involve simultaneous contractions of agonist and antagonist muscles, result in twisting (torsion) postures of the body parts involved, and consistently involve the same pattern of muscle movements.

Dystonia can be classified based upon the age of onset (important as the younger the age of onset, the more likely it is that it will spread to involve other body parts) the body part involved (focal, multifocal, hemidystonia, segmental or generalized) and additionally whether it is:

- primary (usually genetic or undetermined)
- a 'dystonia plus' syndrome with associated parkinsonism or myoclonus
- a secondary dystonia (as a consequence of brain injury or drug therapy)
- heredodegenerative dystonia (associated with diseases such as Huntington chorea)
- dystonia as a feature of other neurological diseases, such as Parkinson disease.

Adult-onset primary focal dystonia is the commonest form and typically affects the neck muscles (cervical dystonia or spasmodic torticollis). Task-specific action dystonias include writer's cramp and musician's dystonia. A comprehensive description of all of the dystonic disorders is beyond the scope of this chapter but important and treatable causes based on the aetiological classification are briefly described below.

DYT1 dystonia (Oppenheim dystonia). This is an autosomal dominant condition with a penetrance of between 30% and 40%. There is a deletion of one of a pair of CAG triplets in the *TOR1A* gene (also known as *DYT1*) on chromosome 9q. This gene codes for torsin A, a heat shock protein. The mutation is commoner in the Ashkenazi Jewish population (1 in 2000). The onset of dystonia is typically before the age of 40 and

it usually involves the upper or lower limbs first. It is a pure dystonic syndrome and conventional imaging (MRI) is normal. Patients with severe generalized dystonia as a consequence of this condition may respond well to deep brain stimulation.

Dopa-responsive dystonia (DRD). Some forms of dystonia respond to levodopa therapy (e.g. dystonia associated with Parkinson disease) but the term 'dopa-responsive dystonia' is specifically reserved for the syndrome arising as a consequence of an autosomal dominant mutation in the GTP-cyclohydrolase 1 gene. It is a disorder of tetrahydrobiopterin synthesis leading to inhibition of phenylalanine, tyrosine and tryptophan hydroxylases and failure of production of catecholamines and indoleamines. Dopa-responsive dystonia should be considered in anyone with early onset dystonia because treatment with L-dopa can lead to dramatic resolution of symptoms. It is an important differential to consider in any patient with presumed dystonic cerebral palsy or young onset spastic paraparesis. The most important diagnostic test is a trial of L-dopa. Other helpful tests include:

- sequence analysis of the GTP-cyclohydrolase 1 gene (positive in approximately 60% of patients with DRD)
- CSF analysis to detect reduced total biopterin and neopterin concentrations
- an oral phenylalanine loading test: a positive result is an elevated plasma phenylalanine:tyrosine ratio following oral phenylalanine.

The gene causing DRD has been mapped to chromosome 14q, although approximately 40–50% of patients with DRD have no identified mutation. If a patient is diagnosed with DRD, there should be an assiduous search for other possibly affected family members (misdiagnosed in some cases to have multiple sclerosis or cerebral palsy), who may benefit from L-dopa therapy.

Tetrahydrobiopterin (TH) deficiency is usually a more serious form of DRD and is transmitted as an autosomal recessive disorder, presenting with dystonia and parkinsonism in infancy. Biochemical analysis of CSF will provide some evidence, but ultimately testing for the gene mutation is required to establish the diagnosis. Decreased CSF concentrations of homovanillic acid and 3-methoxy-4-hydroxyphenylethylene glycol, with normal 5-hydroxyindoleacetic acid concentrations, are the biochemical hallmarks of TH deficiency.

Wilson disease

Wilson disease arises as a consequence of an inborn error of copper metabolism manifesting as hepatic cirrhosis and basal ganglia damage (hepatolenticular degeneration). It can present in a number of different ways and so any patient with a movement disorder under the age of 50 years should be tested. Nearly all patients with neuropsychiatric Wilson disease have Kayser–Fleischer rings (copper deposition at the edge of the corneas), hence slit-lamp examination is an effective screening tool. The average delay from symptom onset to diagnosis is close to 24 months. The initial manifestations of the illness are neurological in ~40% of patients (usually after the age of

12 years). The remainder present with features of liver disease (~40%) or a psychiatric illness.

Wilson disease is inherited as an autosomal recessive trait. The gene responsible lies on chromosome 13q14.3 encoding for a copper-transporting P-type ATPase (ATP7B). The enzyme binds copper in its large N-terminal domain and aids in intracellular processing in the hepatocyte. Intestinal absorption of copper is normal in Wilson disease. There are two pathways for copper excretion from the hepatocyte aided by ATP7B. One is by attachment to caeruloplasmin in the Golgi apparatus, and subsequent delivery of the copper-caeruloplasmin complex into the plasma. A second is promotion of copper excretion into the bile. Mutations of ATP7B lead to failure to excrete copper by both routes, with accumulation of free copper in the circulation and in the tissues, including liver, brain, eyes, kidneys and bones.

The diagnosis and treatment of Wilson disease are discussed in Chapter 14.

Chorea

The term *chorea* is derived from a Greek word meaning 'dance'. Historically, it was first used to describe St Vitus dance, which may have simply described chaotic but voluntary movements occurring in the context of religious fervour. Sydenham chorea describes a childhood chorea occurring as a parainfectious component of acute rheumatic fever but the best known 'choreic' disorder is probably Huntington disease.

Chorea consists of involuntary, continual, abrupt, rapid, brief, unsustained, irregular movements. Its causes can be divided into inherited and acquired. Huntington disease is the commonest inherited disorder: the rest are all rare. Inherited neurometabolic disorders such as the lysosomal storage disorders, amino acid disorders and glucose transporter deficiency syndromes can also have chorea as a predominant phenotypic feature. Of the acquired causes, focal striatal pathology (strokes and space occupying lesions), drug-induced disorders (such as the chorea seen as a complication of long-term dopaminergic therapy in Parkinson disease), chorea gravidarum and thyrotoxicosis are important causes, as are systemic lupus erythematosus, primary antiphospholipid syndrome and central nervous system (CNS) infection.

Myoclonus

The literal meaning of myoclonus is a 'quick movement of a muscle'. Myoclonus can be positive (an active muscle contraction) or negative (sudden and brief loss of muscle tone). The commonest pathological form is probably cortical myoclonus caused by metabolic or toxic encephalopathies. Brief muscle jerks as a consequence of cortical motor discharges, sometimes in response to action or sensory stimuli, are seen.

Asterixis is a form of negative myoclonus in which there is a brief loss of muscle tone, most commonly affecting gait and sustained upper limb posture. 'Liver flap', with an irregular loss of wrist extensor tone seen in hepatic failure and carbon dioxide retention, is a form of metabolic upper limb myoclonus.

Tics

Tics are brief and intermittent movements of muscles and groups of muscles resulting in sudden, abrupt and repetitive movements that may mimic gestures or be fragments of normal behaviour. Tics may therefore be simple or complex movements. Gilles de la Tourette syndrome is the commonest tic disorder and is characterized by motor and vocal or phonic tics beginning in childhood, often accompanied by obsessive-compulsive disorders, poor impulse control and other behavioural problems.

ATAXIA

Ataxia, meaning a 'lack of order', is a term which refers to the type of clumsiness produced by dysfunction of the cerebellum or cerebellar pathways. Impaired balance, gait disorder and clumsiness of the hands, together with dysarthria and jerky tremor of the eyes (nystagmus) are the typical clinical signs. The commonest causes of acquired cerebellar ataxia are acute and chronic alcohol consumption, prescribed drug intoxication (often anti-epileptic drugs), cerebrovascular disease, metastatic cancer and multiple sclerosis. The differential diagnosis in patients where these are excluded is very long and includes all types of neurological pathological processes. The most important feature in the history is the age at onset of ataxia. Ataxia presenting in childhood is more likely to be due to an inherited disorder and onset in adulthood is more likely to be a sporadic ataxia.

The autosomal dominantly inherited cerebellar ataxias (or spinocerebellar ataxias) may be caused by a daunting list of genetic abnormalities, many of which are trinucleotide repeat disorders. No single clinical feature is specific for a particular mutation and even within families there is considerable variability in the phenotype. There are a large number of recessive ataxias but the commonest by far is Friedreich ataxia, which has a prevalence of about 1 in 50 000.

Friedreich ataxia

In this condition, in addition to the ataxia and dysarthria, there may be sensory loss and corticospinal tract signs with absent reflexes. Neurophysiological testing reveals an axonal peripheral neuropathy and skeletal abnormalities (e.g. kyphoscoliosis), hypertrophic cardiomyopathy, and diabetes are common. The genetic abnormality is an expanded trinucleotide repeat of GAA in a gene on chromosome 9q coding for the protein called frataxin. Point mutations are less common. Frataxin is a mitochondrial protein encoded by nuclear DNA and is probably involved in iron transport. Excessive iron accumulation in mitochondria is a feature of the condition.

Ataxia with isolated vitamin E deficiency

This is a rare, but important, disorder that has a phenotype similar to Friedreich ataxia but with additional ophthalmoplegia and retinitis pigmentosa. It arises as

a consequence of defects in the *TTP1* gene coding for α -tocopherol transfer protein that incorporates α -tocopherol into lipoproteins secreted by the liver. The reason for its importance is that it can be treated with vitamin E.

Abetalipoproteinemia

This is a rare autosomal recessive deficiency of apolipoprotein B-containing lipoproteins. The abnormality arises as a consequence of microsomal triglyceride transfer protein (MTP) deficiency. The age of onset is typically in teenage years or early adulthood and the syndrome itself arises as a result of vitamin E deficiency. The main features are a cerebellar ataxia with polyneuropathy, acanthocytosis (characteristically shaped abnormal red blood cells), a coeliac syndrome, and retinal degeneration. It can be treated with high doses of vitamin E.

Ataxia telangiectasia

Ataxia telangiectasia arises as a consequence of a mutation in the 'ataxia telangiectasia mutated' (*ATM*) gene, which codes for a protein kinase that plays an important role in cell cycle control, apoptosis and DNA double-strand break repair. As well as an ataxic neurodegenerative disorder, the abnormality leads to an increased incidence of malignancy. The disorder begins early in childhood, typically at the age of 1–2 years, with truncal ataxia and dysarthria. Oculocutaneous telangiectasias, recurrent sinopulmonary infections and malignancies, particularly leukemia or lymphoma are characteristic features. Plasma α -fetoprotein concentration is elevated, there is evidence of both B cell and T cell immune deficiency and chromosome breaks can be identified.

Early onset ataxia with oculomotor apraxia and hypoalbuminemia

This is caused by a mutation in the aprataxin gene. Cerebellar ataxia and peripheral neuropathy are the commonest features along with oculomotor apraxia. Choreiform movements of the limbs and mental deterioration are also seen. Hypoalbuminemia and hypercholesterolemia are typical and brain MRI or computed tomography show marked cerebellar atrophy. Ataxia with oculomotor apraxia type 2 has also been identified with a similar clinical picture but with mutations in the senataxin gene.

Fragile X-associated tremor/ataxia syndrome

This condition is associated with a tremor that appears very similar to that of essential tremor. Other clinical features include a cerebellar ataxia and mild to moderate executive dysfunction. Both sexes can be affected. The disease is rare and is characterized pathologically

by intranuclear inclusions. An increased T2 signal in the middle cerebellar peduncle is a relatively specific MRI appearance.

Hexosaminidase deficiency (GM2 gangliosidosis)

This group of conditions is characterized by accumulation of GM2 ganglioside and glycosphingolipids within neuronal cells (and other cell types) owing to a genetically determined deficiency of the enzyme hexosaminidase. Depending on subunit composition, there are three possible isoenzymes: hexosaminidase A ($\alpha\beta$), hexosaminidase B ($\beta 2$) and hexosaminidase S ($\alpha 2$). The two classic subtypes are Tay–Sachs disease and Sandhoff disease.

In Tay–Sachs disease, there is absence of hexosaminidase A ($\alpha\beta$) and hexosaminidase S ($\alpha 2$). Hexosaminidase B ($\beta 2$) activities are normal or slightly elevated. The abnormality lies in the α -subunit, possibly owing to a mutation at the locus on chromosome 15. In Sandhoff disease, there are low activities of both hexosaminidase A and B. Hexosaminidase S activity is increased. There is a deficiency of the β -subunit owing to a mutation at the locus on chromosome 5.

There is a spectrum of clinical presentations of these conditions:

- *infantile onset* is the best known and presents at the age of six months with myoclonic seizures and progressive visual loss. A macular cherry-red spot is seen. There is progressive dementia and death within 4–6 years. This group includes both Tay–Sachs and Sandhoff disease
- *juvenile onset* is at 1–4 years, with dementia, seizures and ataxia. A cherry-red spot at the macula may or may not be present
- *an ataxic form* of GM2 gangliosidosis is characterized by a slowly progressive cerebellar or spinocerebellar ataxia with onset in late childhood or the adolescent period, which can be the manifestation of either hexosaminidase A or hexosaminidase A and B deficiencies
- *a spinal muscular atrophy form* is recognized with slowly progressive generalized muscle wasting and weakness. Only a deficiency of hexosaminidase A has been identified in this form, which can present in childhood or early adulthood
- *an adult onset* type is characterized by early onset of dementia, seizures and normal pressure hydrocephalus.

In addition to measuring hexosaminidase activities in plasma, leukocytes or cultured skin fibroblasts, a rectal biopsy should be done in adult onset cases to look for the characteristic changes in autonomic nerve fibres that are apparent on electron microscopy.

Cerebrotendinous xanthomatosis (cholestanolosis)

This is a leukodystrophy characterized by severe progressive ataxia and limb spasticity by the age of 30.

Characteristic appearances include tendon xanthomas and zonular cataracts. Dementia may occur at a young age before the other physical signs are evident and it can be difficult to make a clinical diagnosis at this early stage. Pathologically, there is widespread CNS demyelination with degeneration of the spinal cord. See page 695 for further details about the underlying biochemical abnormalities and potential treatment.

Neuronal ceroid lipofuscinosis

This is another condition in which the age of presentation determines the subtype. Early development is usually normal, with the age of onset from six months to 30 years. Patients present with myoclonus, ataxia and visual loss (except in the late onset variety). If seizures are a feature, they tend to be refractory to treatment. Most cases are autosomal recessive but the adult onset can be autosomal dominant. The incidence varies throughout the world with a particular predilection for the infantile subgroup in Finland of 1:13 000 live births.

Histological examination reveals an accumulation of autofluorescent lipopigment inclusions in body tissues and the diagnosis is made from biopsy of skin, conjunctival or rectal tissue. Abnormal accumulation of S-methylated methionine has been found in the brains of affected infants and of ϵ -N-trimethyl lysine in those from patients with the juvenile onset form. There is a profound deficiency of endo- β -N-acetylglucosaminidase, which is required for the cleaving of chilibiosyl residues. All forms of this condition lead to learning disability, seizures, ataxia and, in the early onset types, visual loss. Life expectancy is severely limited.

Coeliac disease

This is an immunologically mediated disease (see also Chapter 12) strongly associated with certain human leukocyte antigens. It is common and in large case series, neurological complications have been reported in up to 6% of patients. Cerebellar ataxia is one of the commonest neurological associations. The occurrence of anti gliadin antibodies in otherwise idiopathic ataxia requires further dedicated research studies and it remains possible that adult onset sporadic ataxia with no clear cause and associated coeliac disease is no more than a chance occurrence.

INFLAMMATORY DISORDERS OF THE CENTRAL NERVOUS SYSTEM

Inflammation is a common pathological process within the CNS, by far the commonest form being multiple sclerosis (MS). Diagnosis is made on clinical grounds together with MRI evidence. Oligoclonal bands are present in the CNS in up to 95% of patients with clinically definite MS (see Chapter 34). Other inflammatory conditions affecting the CNS include systemic lupus erythematosus, Sjögren disease, Behçet disease, polyarteritis nodosa, paraneoplastic disorders, acute disseminated encephalomyelitis and vasculitis. Mimics of MS also

include granulomatous disorders (sarcoidosis, Wegener granulomatosis), diseases of myelin (metachromatic leukodystrophy, adrenomyeloleukodystrophy), vitamin B₁₂ deficiency, and certain infectious diseases (Lyme disease, neuroborreliosis, human T-lymphotropic virus type 1, HIV, progressive multifocal leukoencephalopathy and neurosyphilis).

CONCLUSION

The conditions discussed above represent a very small fraction of all neurological disorders. There are many more for which clinical biochemistry, at our current level of understanding, has little to offer in everyday clinical practice. These include headache, epilepsy, sleep disorders, cerebrovascular disease (beyond glucose and lipid chemistry), neurological malignancies, dementia and neurological trauma. There are also many conditions, for which biochemical perturbations are an important part of diagnoses, however a full discussion is beyond the scope of this chapter. These include many other inherited metabolic diseases, a multitude of which have neurological manifestations.

ACKNOWLEDGEMENT

We would like to acknowledge the contribution of J. Gareth Llewelyn, Mark Cossburn, Alistair Church and Huw R. Morris, who wrote the chapter for the previous edition of the book.

Further reading

- Albers JW, Fink JK. Porphyric neuropathy. *Muscle Nerve* 2004;30:410–22.
- Daroff RB, Fenichel GM, Jankovic J et al. editors. *Bradley's neurology in clinical practice*. 6th ed Philadelphia: Saunders; 2012.
- Donaldson I, Marsden CD, Schneider S et al. *Marsden's book of movement disorders*. Oxford: Oxford University Press; 2012.
- This book represents the final work of the late Professor C. David Marsden, who was the most influential figure in the field of movement disorders, in terms of his contributions to both research and clinical practice, in the modern era.*
- Hauser AC, Lorenz M, Sunder-Plassmann G. The expanding clinical spectrum of Anderson-Fabry disease: a challenge to diagnosis in the novel era of enzyme replacement therapy. *J Intern Med* 2004;255:629–36.
- Hughes R. Rational testing: investigation of peripheral neuropathy. *Br Med J* 2010;341:c6100.
- Mehta A, Beck M, Sunder-Plassmann G, editors. *Fabry disease: perspectives from 5 years of FOS*. Oxford: Oxford PharmaGenesis; 2006.
- Pop-Busui R, Sima A, Stevens M. Diabetic neuropathy and oxidative stress. *Diabetes Metab Res Rev* 2006;22:257–73.
- Sullivan KA, Feldman EL. New developments in diabetic neuropathy. *Curr Opin Neurol* 2005;18:586–90.
- Surtees R. Biochemical pathogenesis of subacute combined degeneration of the spinal cord and brain. *J Inher Metab Dis* 1993;16:762–70.

Internet resources

- Washington Neuromuscular Disease Center: neuromuscular.wustl.edu; [Accessed March 2013].
- Website of the Washington Neuromuscular Disease Center.*
- American Academy of Neurology. www.aan.com/go/practice/guidelines; [Accessed March 2013].
- The clinical practice guidelines page of the American Academy of Neurology.*

Lipids and disorders of lipoprotein metabolism

Graham R. Bayly

CHAPTER OUTLINE

INTRODUCTION 703

LIPIDS 703

- Sterols 703
- Fatty acids 704
- Triglycerides 704
- Phospholipids 704
- Eicosanoids 705
- Sphingolipids 705
- Nuclear lipids 706

LIPOPROTEINS 706

- Chylomicrons 707
- Very low density lipoproteins 708
- Intermediate density lipoproteins 708
- Low density lipoproteins 708
- High density lipoproteins 708
- Lipoprotein(a) 708
- Lipoprotein X 708

APOLIPOPROTEINS 708

- Apolipoprotein A 709
- Apolipoprotein B 709
- Apolipoprotein C 709
- Apolipoprotein D 710
- Apolipoprotein E 710
- Apolipoprotein M 710
- Apolipoprotein(a) 710

CHOLESTEROL ABSORPTION 710

TRIGLYCERIDE DIGESTION 711

BILE ACID METABOLISM 711

LIPOPROTEIN METABOLISM 711

- Assembly of apolipoprotein B-containing lipoproteins 711
- Exogenous pathway 713
- Endogenous pathway 713
- High density lipoprotein metabolism 714

ENZYMES INVOLVED IN LIPOPROTEIN METABOLISM 716

- Lecithin cholesterol acyltransferase 716
- Lipases 716
- Acyl-CoA:cholesterol acyltransferase 718

TRANSFER PROTEINS INVOLVED IN LIPOPROTEIN METABOLISM 718

- Cholesteryl ester transfer protein (CETP) 718
- Phospholipid transfer protein (PTP) 718
- Fatty acid transport proteins 718

RECEPTORS INVOLVED IN LIPOPROTEIN METABOLISM 718

- The LDL receptor 718
- LDL receptor-related protein 719
- Scavenger receptor class B type 1 719
- Other scavenger receptors 719
- Peroxisome proliferator-activated receptor family 720
- Other nuclear receptors 720

OTHER PROTEINS INVOLVED IN LIPOPROTEIN SYNTHESIS, TRANSPORT AND METABOLISM 720

- Microsomal triglyceride transfer protein 720
- ATP binding cassette transporter family 720
- Proprotein convertase subtilisin kexin 9 720
- Sterol regulatory element binding proteins 721
- Sortilins 721
- Glycosylphosphatidylinositol-anchored HDL-binding protein 1 721
- Angiopoietin-like protein 3 721

CLASSIFICATION OF LIPOPROTEIN DISORDERS 721

THE PRIMARY DYSLIPOPROTEINAEMIAS 723

- Hypobetalipoproteinaemia 723
- Familial combined hyperlipidaemia 724
- Familial hypertriglyceridaemia 724
- Remnant hyperlipoproteinaemia 725
- Familial hypercholesterolaemia 726
- Polygenic hypercholesterolaemia 728
- Dysalphalipoproteinaemias 728
- Disorders of HDL metabolism 728

ACQUIRED HYPERLIPIDAEMIAS 729

- Diabetes mellitus 729
- Hypothyroidism 730
- Nephrotic syndrome 730
- Chronic kidney disease 730
- Renal transplantation 731

Liver disease 731
 Alcohol 731
 Drug-related hyperlipidaemia 731
ACQUIRED HYPOLIPIDAEMIA 732
INVESTIGATION OF LIPID DISORDERS 732
 Total cholesterol 732
 Triglycerides 732
 High density lipoprotein cholesterol 732

Low density lipoprotein cholesterol 732
 Apolipoproteins 733
 Post-heparin lipolytic activity 733
 Lipoprotein separation techniques 733
 Genotyping 734
TREATMENT OF HYPERLIPIDAEMIA 734
CONCLUSION 736

INTRODUCTION

'Lipid' is the term used to describe a number of substances of diverse chemical structure that bear little functional relationship to each other but which have in common the property of being soluble in organic solvents and virtually insoluble in water. Lipoproteins are macromolecular protein complexes that allow hydrophobic lipids to be transported within the hydrophilic environment of the circulation. Lipids are essential for health, but excessive concentrations of cholesterol and triglycerides in the circulation, whether due to lifestyle factors or to inherited disorders of lipoprotein metabolism, are major factors in the development of atherosclerosis and cardiovascular disease. These conditions are the focus of this chapter, although other lipids and their functions are discussed briefly.

LIPIDS

Lipids can be broadly divided into sterols, including cholesterol; fatty acids or substances containing fatty acids such as triglycerides and phospholipids; eicosanoids; the fat-soluble vitamins (A, D, E and K), and sphingolipids. The major classes of lipids and their principal functions are summarized in Table 37.1.

TABLE 37.1 Major classes of lipids and their function

Lipid	Function
Cholesterol	Structural component of membranes; precursor for bile acid and steroid synthesis
Fatty acids	Energy source
Triglycerides	Energy store
Phospholipids	Structural component of membranes
Eicosanoids	Multiple, including effects on blood coagulation, bronchial and vascular contractility, reproduction
Sphingolipids	Central nervous system; blood group substances
Fat-soluble vitamins	
Vitamin A	Vision
Vitamin D	Calcium homeostasis and maintenance of bone integrity; various other functions, e.g. in immunomaturation
Vitamin E	Neural function; antioxidant
Vitamin K	Activation of clotting factors

Sterols

Cholesterol

Cholesterol (Fig. 37.1) is the major sterol in humans, being present in all body cells and most body fluids. The majority of the cholesterol in the body is in the free, unesterified form; it is this form that is the structural component of cell membranes. Cholesteryl esters in normal cells represent a store for future use and appear microscopically as intracellular droplets. Cholesterol is present in the diet but most cholesterol in the body is made by de novo synthesis from acetate. The rate-limiting step in the synthetic pathway is the conversion of 3-hydroxy-3-methylglutaryl-coenzyme A (HMG-CoA) to mevalonate, catalysed by the enzyme HMG-CoA reductase. The liver is responsible for most cholesterol synthesis. Cholesterol is a precursor for the synthesis of gonadal and adrenal steroid hormones, vitamin D and bile acids.

Although only a small amount of the body's cholesterol pool comes from dietary cholesterol, this has an important role in regulating the rate of cholesterol synthesis. The liver is the key organ in maintaining cholesterol balance; any excess cholesterol is excreted by the liver into the bile, either directly, or after conversion into bile acids.

Cholesterol and membranes. Cholesterol is a major component of cell membranes. Cholesterol and sphingomyelin form plasma membrane 'lipid rafts' or caveolae. Caveolae are cell surface invaginations found in differentiated cells and characterized by the presence of a protein, caveolin-1; they are sites where signalling molecules are concentrated. In order for these signalling molecules to function, the cholesterol concentration of the plasma membrane must remain constant. This is achieved by a regulatory system that senses the cholesterol content of

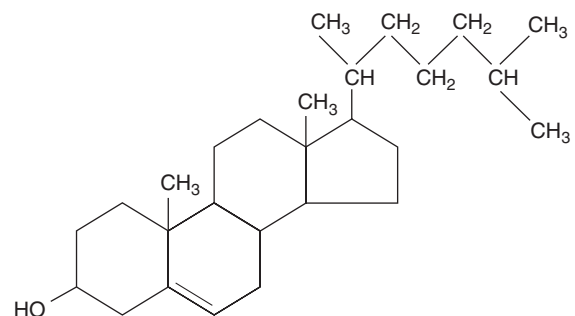


FIGURE 37.1 ■ Structure of cholesterol.

the membrane and modulates the transcription of genes encoding proteins involved in cholesterol synthesis (e.g. HMG-CoA synthase and HMG-CoA reductase) and cholesterol uptake (e.g. the LDL receptor). The system that does this is a family of membrane-bound transcription factors called sterol regulatory element-binding proteins (see p. 721).

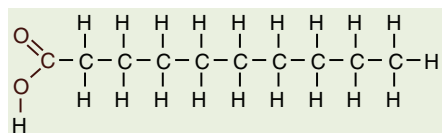
Phytosterols

Phytosterols are sterols derived by plants; they differ slightly from cholesterol. There are two general types: δ^3 -phytosterols (e.g. β -sitosterol) and 5α -reduced phytosterols, otherwise referred to as 'stanols'. In natural foods, the sterols predominate. Both sterols and stanols are incorporated into commercially available foods promoted as cholesterol-lowering agents: they act by competing with cholesterol for absorption from the gut, and may lower plasma cholesterol concentration by 10–15%.

Fatty acids

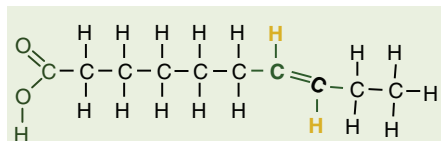
Fatty acids (Fig. 37.2) have the general chemical formula RCOOH. Those relevant to human nutrition are the long chain (C_{12} – C_{20}) fatty acids containing even numbers of carbon atoms. They are further defined as saturated, for example stearic (C18:0); monounsaturated, for example oleic (C18:1), and polyunsaturated, for example linoleic (C18:2) and linolenic (C18:3), the second figures indicating the number of double bonds. In general, dietary saturated fatty acids originate from animals and the unsaturated fatty acids from plants. There are, however, exceptions, for example palmitic acid (C16:0) is saturated but derived from palm oil, and the ω -3 series, which are unsaturated but found in fish. The position of the carbon of the first double bond in the polyunsaturated fatty acids differentiates the ω -6 series (where the first double bond starts beyond the 6th carbon atom from

Saturated



Monounsaturated

Trans



Cis

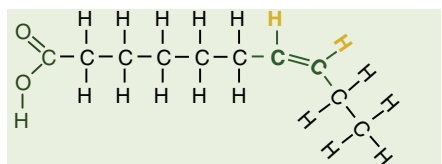


FIGURE 37.2 ■ Structure of fatty acids.

the methyl end of the molecule), from the ω -3 series, where it starts beyond the 3rd carbon atom.

The long chain fatty acids are oxidized for energy production by a process known as β -oxidation; this results in the sequential shortening of the chain by two carbon atoms and the production of acetyl-CoA. Small amounts of long chain fatty acids are elongated to very long chain fatty acids (VLCFAs), which have a structural function in certain specialized cells.

Triglycerides

Triglycerides (see Fig. 37.3) comprise three fatty acids esterified with a glycerol backbone. 'Triacylglycerols' is the correct chemical name but they are more commonly known as 'triglycerides' and this term will be used throughout this chapter. Triglycerides are the major dietary fat. They are hydrolysed in the gut by lipases to fatty acids and monoglycerides. The monoglycerides undergo re-esterification in enterocytes and subsequent incorporation into chylomicrons. The major sites of endogenous triglyceride synthesis are the liver and adipose tissue. In normal circumstances, hepatic triglyceride is secreted in very low density lipoproteins (VLDL). In certain pathological states, triglyceride accumulates in hepatocytes, leading to hepatic steatosis. Adipose tissue triglyceride represents the major energy store of the body. Fatty acids are mobilized from adipose tissue triglycerides by the action of hormone-sensitive lipase (HSL), which is activated by glucagon and adrenaline (epinephrine) and inhibited by insulin.

Phospholipids

Phospholipids, like triglycerides, have a glycerol backbone; in phospholipids, this is esterified with two fatty acids and the third hydroxyl group is linked via a phosphodiester bond to an amino alcohol such as choline, serine or ethanolamine (Fig. 37.4). Phospholipids are therefore amphipathic molecules, with both hydrophilic (phosphate group) and hydrophobic (fatty acid) domains. This property is responsible for the capacity of phospholipids to solubilize other lipids and accounts for their location on the surfaces of lipoprotein molecules and in the cell membrane lipid bilayer.

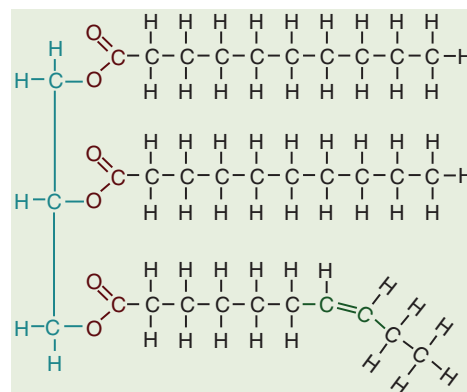


FIGURE 37.3 ■ Structure of triglycerides.

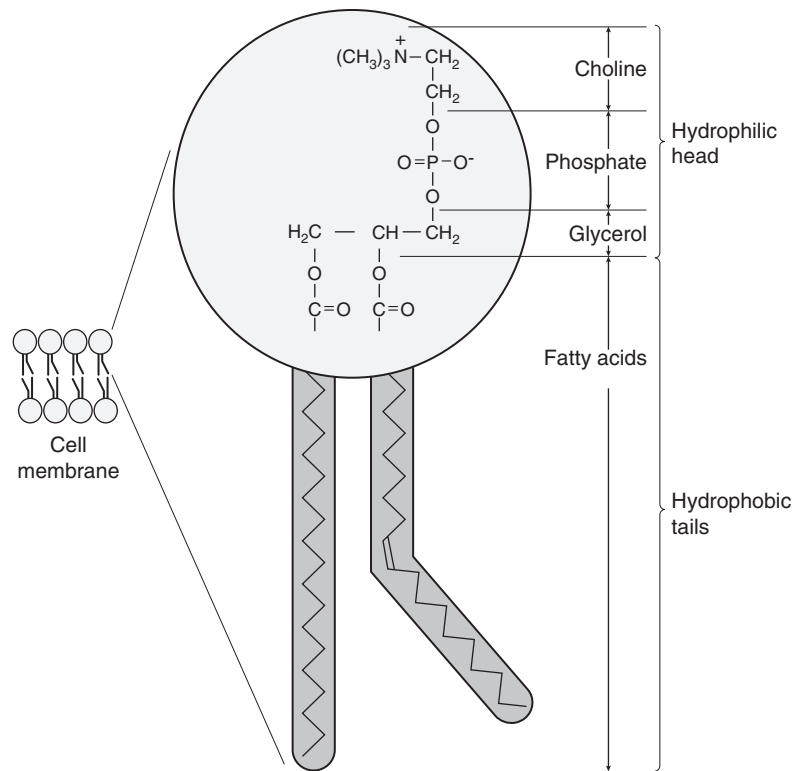


FIGURE 37.4 ■ Structure of phospholipids.

Eicosanoids

This group of compounds takes its name from the systematic name of C_{20} (eicosa-) fatty acids from which they are derived. It includes prostaglandins, thromboxanes and leukotrienes. These were originally named because they were found in prostate, platelets (thrombocytes) and white cells (leukocytes), respectively. They have major effects on the immune response, reproductive function (including the induction of labour), cholesterol metabolism, smooth muscle function (causing vasoconstriction or dilatation), platelet aggregation and thrombosis.

Prostaglandins contain a substituted cyclopentane ring. Thromboxanes are oxygenated eicosanoids closely related to the prostaglandins. The rate-limiting step in the synthesis of both (see Fig. 37.5) is the phospholipase A_2 -mediated release of fatty acids from phospholipids. The principal precursor of prostaglandins is arachidonic acid ($\text{C}_{20:4}$, eicosatetraenoic acid). The first two steps in their synthesis are catalysed by prostaglandin endoperoxide synthase, which has both cyclooxygenase and peroxidase activity. The major products of this enzyme's activity are prostaglandin G_2 (PGG_2) and prostaglandin H_2 (PGH_2). These are both prostaglandins of the 2 series, having 2 carbon-carbon double bonds. Prostaglandin H_2 is converted to thromboxane A_2 (TXA_2) in platelets and prostacyclin (PGL_2) in the arterial wall. Arachidonic acid also serves as a substrate for various lipoxygenases, which generate the leukotrienes, inflammatory mediators and potent stimulators of muscle contraction. Eicosatrienoic acid ($20:3$) and eicosapentenoic acid ($20:5$), are alternative precursors for prostaglandin synthesis. The latter generates prostaglandins of the 3 series and leukotrienes of the 5 series.

Eicosanoid imbalance is of interest in atherosclerosis because several recognized risk factors for atherosclerosis, such as smoking, hypertension and diabetes mellitus, are associated with changes in eicosanoid production.

Aspirin inhibits cyclooxygenase, thereby reducing prostaglandin, especially platelet TXA_2 , synthesis. This is the rationale for the use of aspirin in the prophylaxis of atherosclerotic vascular disease. Steroids directly inhibit the release of arachidonic acid from membrane phospholipids.

Sphingolipids

The backbone of sphingolipids is sphingosine or a similar long chain base. The sphingolipids vary in chain length from C_{14} to C_{20} . Sphingosine itself has 18 carbon atoms and is formed from the condensation of palmitoyl CoA ($\text{C}_{16:0}$) with serine. Another fatty acid chain is joined to sphingosine through an amide link to form a ceramide. Ceramides have a free hydroxyl group that allows reaction with another component. If this contains a phosphate group, then the resultant products are a type of phospholipid called sphingophospholipids, also termed sphingomyelins (Fig. 37.6), which are essential components of nerve cells. If a carbohydrate group is attached to ceramide then glycosphingolipids, or simply glycolipids, are formed. These include cerebrosides, sulfatides, globosides and gangliosides. Glycosides of ceramide are referred to as cerebrosides: they are present in relatively high concentrations in brain. Many glycosphingolipids contain oligosaccharides; those that contain more than one molecule of sialic acid are referred to as gangliosides.

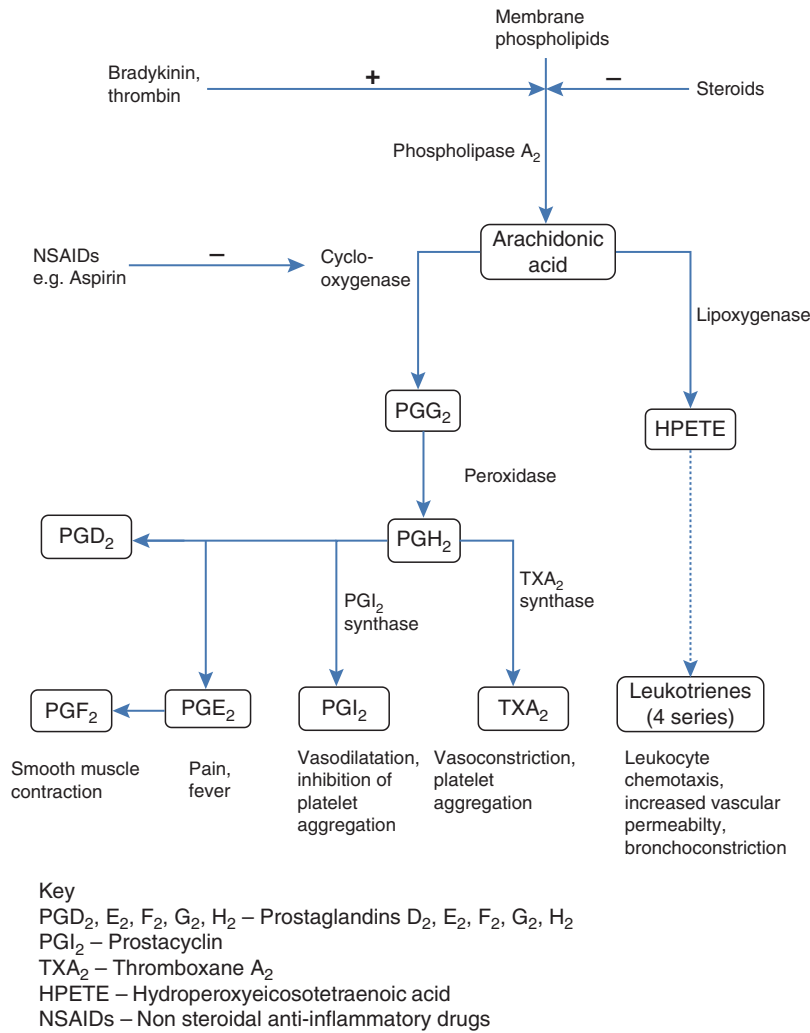


FIGURE 37.5 ■ An outline of eicosanoid synthesis.

Nuclear lipids

The nucleus is a highly structured organelle. For a long time, it was thought that lipids, which quantitatively are a minor component of the nucleus, only served a structural role and originated in the cytoplasm. It is now recognized that lipids serve other functions including signalling and modulating. Lipid-metabolizing enzymes have been demonstrated to be present within the nucleus. Phospholipids are the predominant class of lipid in the nucleus, with lesser amounts of cholesterol, free fatty acids, diglycerides and sphingolipids. The nucleus is surrounded by the nuclear envelope, which comprises an outer nuclear membrane and an inner nuclear membrane. Cholesterol has a structural role in the outer nuclear membrane, which is continuous with the endoplasmic reticulum; the inner nuclear membrane is associated with the nuclear lamina and chromatin and is deficient in cholesterol. The inner and outer nuclear membranes are joined by the pore membranes at the nuclear pores, which are associated with the nuclear pore complexes that allow passive transfer of molecules <50 kDa between the cytoplasm and nucleoplasm. The passage of larger molecules is energy dependent and requires a nuclear localization signal.

Lipids with very long chain fatty acids are associated with the pore membrane–nuclear pore complexes and appear to be essential for maintaining their function.

LIPOPROTEINS

The lipoproteins are submicroscopic, macromolecular complexes of lipids (cholesterol, triglycerides, phospholipids) and proteins (apolipoproteins, enzymes), held by non-covalent forces. The basic structure of lipoproteins is a hydrophobic core of triglycerides and/or cholesteryl esters surrounded by a layer of amphipathic phospholipids, unesterified cholesterol and proteins (see Fig. 37.7). The hydrophilic surface protects the hydrophobic core from the aqueous environment. Lipoproteins differ in their relative concentrations of protein to lipid and in their constituent lipids and proteins (Table 37.2). The densities of lipoproteins are inversely related to their size. The lipoproteins can be classified on the basis of their size, density or protein composition. The nomenclature of the lipoproteins is based on their density: chylomicrons (<0.95 g/mL); VLDL (0.95–1.006 g/mL); intermediate density lipoproteins (IDL) (1.006–1.019 g/mL); low density

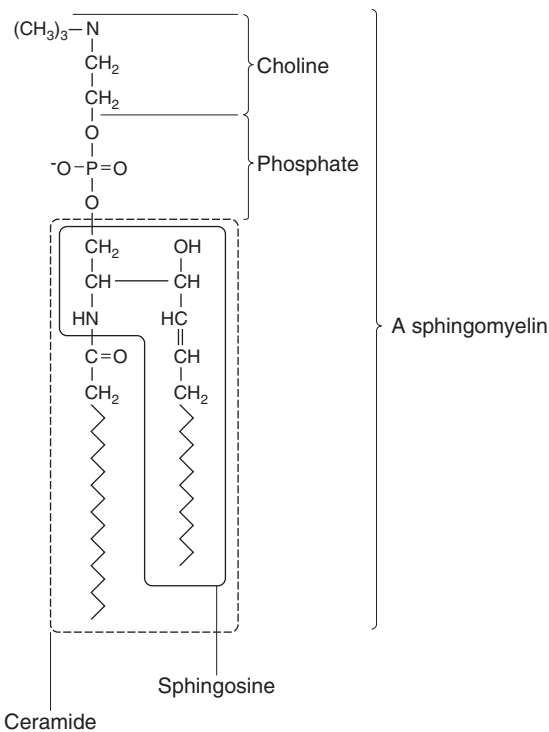


FIGURE 37.6 ■ Structure of sphingolipids.

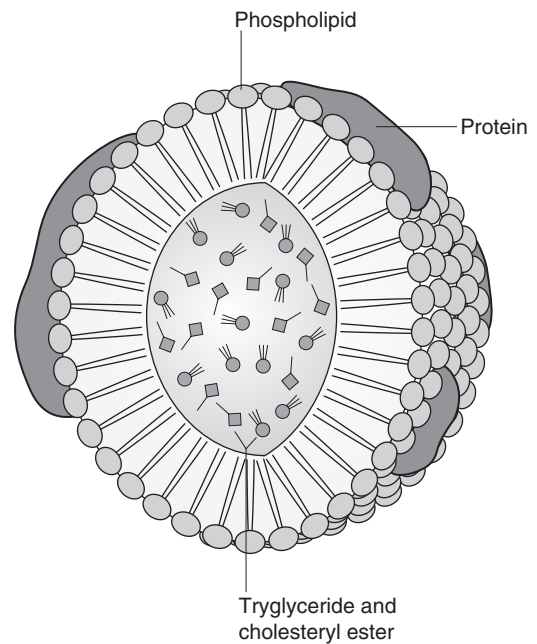


FIGURE 37.7 ■ Generic lipoprotein structure.

TABLE 37.2 Characteristics and major functions of apolipoproteins

Name	Chromosome	Amino acids	Mol wt (kDa)	Structural role	Function
A-I	11	243	29	HDL	LCAT activator
A-II	1	77	17 (dimer)	HDL	LPL regulator, LCAT and CETP cofactor
A-IV	11	396	44	Chylomicrons, HDL	LCAT activator
A-V	11	366	41	Chylomicrons, VLDL, HDL	LPL activator, chylomicron assembly
C-I	19	57	7	Chylomicrons, VLDL, HDL	LCAT activator
C-II	19	79	9	Chylomicrons, VLDL, HDL	LPL activator
C-III	11	79	9	VLDL, chylomicrons, HDL	LPL inhibitor, VLDL assembly
B-100	2	4536	500	VLDL/IDL/LDL	LDL-R ligand
B-48	2	2152	240	Chylomicrons and remnant particles	Structural component of chylomicrons
D	3	169	33		Transport of small lipophilic molecules
E	19	299	34	Chylomicrons, VLDL, remnants	LDL-R ligand for chylomicron remnants; LRP ligand
M	6	188	26		Transport of small lipophilic molecules
(a)	6	Variable	200–800	Lp(a)	

lipoproteins (LDL) (1.019–1.063 g/mL) and high density lipoproteins (HDL) (1.063–1.210 g/mL). The classes are not homogeneous; each represents a continuum of particles of differing size, density and fate and, in the case of VLDL, also of origin. The physicochemical characteristics of the principal lipoproteins are summarized in Table 37.3.

The apolipoprotein (apo) B-containing lipoproteins contain only one molecule of apo B per lipoprotein particle whereas multiple molecules of the other apolipoproteins are present in other lipoprotein particles.

Chylomicrons

Chylomicrons are the largest and most buoyant class of lipoprotein. The major protein component is apo B-48 but they also contain apo A-I, apo A-II and apo A-IV. After secretion, they acquire apo E and apo C from HDL. Chylomicrons are formed in the intestine and are the transport vehicle for dietary fat. The largest chylomicron particles have a diameter of over 1000 nm, whereas the smallest (75–200 nm) overlap with the apo B-100-containing lipoproteins. Some of the smaller particles within the chylomicron range are this size when they are

TABLE 37.3 Characteristics of the major lipoprotein classes

Lipoprotein	Hydrated density (g/mL)	Diameter (nm)	Molecular weight (kDa)	Electrophoretic mobility	Major apolipoprotein
Chylomicron	<0.95	75–1200	50 000–1 000 000	Origin	B-48
VLDL	0.95–1.006	30–200	10 000–80 000	Pre- β	B-100
IDL	1.006–1.019	25–35	5000–10 000	Slow pre- β	B-100
LDL	1.019–1.063	18–25	2300	β	B-100
HDL ₂	1.063–1.12	9–13	400	α	A-I (and A-II)
HDL ₃	1.12–1.21	7–9	200	α	A-I
Pre- β ₁ HDL	1.21	5–7	70	Pre- β	A-I

secreted by enterocytes, while others represent partially delipidated 'remnant' particles. The core of chylomicrons is composed predominantly of triglycerides derived from the diet.

Very low density lipoproteins

These are the largest of the lipoproteins containing endogenously produced lipids. The major protein component of VLDL is apo B-100 but they also contain apo C-I, apo C-II, apo C-III, apo E and small amounts of apo A. Like chylomicrons, VLDLs acquire the majority of their component apo E and apo C from HDL in the circulation; the core of VLDLs is composed predominantly of triglycerides. In contrast to chylomicrons, the triglycerides in VLDL are endogenous in origin.

Intermediate density lipoproteins

These particles are produced during the conversion of VLDL to LDL; their densities lie between those of these lipoproteins. The core of IDLs contains cholesteryl esters and triglycerides.

Low density lipoproteins

These are the major cholesterol-containing lipoproteins and represent the end-product of VLDL catabolism. The core of LDLs comprises mainly cholesteryl esters; the protein component is apo B-100.

High density lipoproteins

These are the smallest and densest of the lipoproteins. They may be sub-classified on the basis of size, density, shape, surface charge and electrophoretic mobility, as well as apolipoprotein composition (Table 37.3). High density lipoprotein is usually divided into three major subclasses. Nascent, discoidal or pre- β ₁ HDL comprises predominantly apo A-I and phospholipid. It is the preferred substrate for the ATP binding cassette transporter A1 (ABCA1), which actively exports free cholesterol from peripheral cells and macrophages. HDL₃ is formed from pre- β ₁ HDL by the acquisition of free cholesterol. It is the preferred substrate for lecithin cholesterol acyl transferase (LCAT), which esterifies free cholesterol, increasing the size of the particle and allowing the uptake of more free cholesterol, producing the larger and more cholesterol-rich HDL₂.

High density lipoprotein particles may contain apo A-I alone (LpA-I), both apo A-I and apo A-II (LpA-I/A-II) or apo A-II alone (LpA-II). LpA-I predominates in HDL₂, whereas LpA-I/A-II predominate in HDL₃. LpA-II represents a very small proportion of both HDL₂ and HDL₃.

Lipoprotein(a)

Lipoprotein(a) (Lp(a)) consists of LDL with its apo B-100 bound by a disulfide bond to apolipoprotein(a) (apo(a)). It is thought to be assembled extracellularly, either in the circulation or on the surfaces of hepatocytes.

The plasma concentration of Lp(a) is genetically determined and is inversely related to the length of the apo(a), so that the greater the chain length, the lower the concentration. Epidemiological studies suggest that a high Lp(a) concentration is an independent risk factor for cardiovascular disease (CVD), particularly in subjects with familial hypercholesterolaemia.

Lipoprotein X

Lipoprotein X is a lipoprotein that is found only in the plasma of subjects with cholestasis or who have familial lecithin cholesterol acyltransferase deficiency. It is composed of phospholipids, free cholesterol and proteins; the major protein is albumin but small amounts of apo C and apo D are also present. It contains no apo B. Unlike all other lipoproteins, it migrates towards the cathode on agarose gel electrophoresis.

APOLIPOPROTEINS

The apolipoproteins are amphipathic; their hydrophobic regions interact with the lipids in the lipoprotein particle while their hydrophilic regions allow interaction with the aqueous environment. They have three functions: they provide the structural element to the lipoprotein particles, they act as ligands for specific receptors and they also act as activators or inhibitors of specific enzymes involved in lipoprotein metabolism.

On the basis of electrophoretic mobility, HDL and LDL were originally referred to as α - and β -lipoproteins. The nomenclature of the corresponding apolipoproteins has arisen from this, apo A being the apolipoprotein derived from HDL (α -lipoprotein) and apo B being derived from LDL (β -lipoprotein).

Apolipoprotein A

Apolipoprotein A-I

Apolipoprotein A-I (apo A-I) (molecular weight 29 kDa) is the major protein of HDL, constituting 70–80% of HDL protein. It is synthesized primarily in the liver and small intestine. In addition to its structural role in HDL, it is also an activator of lecithin cholesterol acyltransferase (LCAT). Reverse cholesterol transport is dependent on the ability of apo A-I to promote cellular cholesterol efflux, to bind to lipids, to activate LCAT and, within mature HDL, to interact with lipid transfer proteins and specific receptors. The gene for apo A-I (*APOA1*) is part of a gene cluster on the long arm of chromosome 11 that includes *APOC3*, *APOA4* and *APOA5*.

Epidemiological studies have shown that plasma apo A-I concentrations, like those of HDL-cholesterol (HDL-C), are inversely related to cardiovascular risk.

Apolipoprotein A-II

Apolipoprotein A-II (apo A-II) (molecular weight 17 kDa, as homodimer) is also synthesized in the liver and, to a lesser extent, the small intestine. It accounts for about 20% of HDL protein. Some HDL contains apo A-I and apo A-II while some HDL contains apo A-I alone. A small amount of plasma apo A-II is associated with chylomicrons and VLDL. Apo A-II regulates lipoprotein lipase (LPL) activity, and is a cofactor for LCAT and for cholesteryl ester transfer protein (CETP). Like apo A-I, it appears to be inversely related to the risk of coronary disease. It may play a role in the remodelling of HDL, possibly by an effect on the reactivity of HDL towards lipid transfer proteins, enzymes and receptors, including the scavenger receptor B1 (SRB1) (see p. 719).

Apolipoprotein A-IV

Apolipoprotein A-IV (apo A-IV) (molecular weight 44 kDa) is synthesized only in the small intestine. It has been suggested that it may have a role in intestinal lipid transport, increasing the residence time of nascent chylomicron particles, allowing for greater expansion of their cores and thus capacity to transport triglycerides. The majority of apo A-IV in plasma exists in the free form. A small amount is associated with HDL and chylomicrons. In vitro, apo A-IV activates LCAT, although not as effectively as does apo A-I. It may also be necessary for the maximal activation of LPL by apo C-II. Over-expression of *APOA4* in mice results in increased plasma concentrations of total- and HDL-cholesterol, and triglycerides; despite this, it protects against diet-induced atherosclerosis. In humans, apo A-IV deficiency has been reported in patients with apo A-I and apo C-III deficiency, and this may account for the fat malabsorption seen in affected individuals.

Apolipoprotein A-V

The *APOA5* gene is expressed in liver, and apolipoprotein A-V (apo A-V), in contrast to other apolipoproteins, is present at very low concentrations in the plasma (approximately 5 nmol/L). It is found primarily in HDL.

Apo A-V affects plasma triglycerides through an effect on the lipolysis of the triglyceride-rich lipoproteins, possibly by binding to the lipoprotein, endothelial proteoglycans and LPL and thus stabilizing the lipolytic machinery. Although plasma concentrations of apo A-V show little correlation with plasma triglyceride concentration or with prevalence of cardiovascular disease, genetic studies have shown polymorphisms in *APOA5* to be strong determinants of both. Genetic variants in humans have been identified, in association with both high and low triglyceride concentrations. Deficiency leads to reduced LPL activity and a type V dyslipidaemia (p. 726).

Apolipoprotein B

This lipoprotein exists in two forms; apolipoprotein B-100 (apo B-100), which is made in the liver and is the structural protein of VLDL, IDL and LDL, and apo B-48, which is synthesized in the intestine and is incorporated into chylomicrons. Both apo B molecules remain with the lipoprotein particle in which they are secreted throughout the lifespan of that particle, unlike the other apolipoproteins, which readily transfer between different classes of lipoproteins. Increased plasma concentrations of apo B-containing lipoproteins confer an increased risk for the development of atheroma. Both forms of apo B are produced from the *APOB* gene; post-transcriptional editing of the mRNA in the intestine leads to the production of apo B-48.

Apolipoprotein B-100

Apolipoprotein B-100 (apo B-100) (molecular weight 500 kDa) is necessary for the assembly and secretion of VLDL. It contains several very hydrophobic areas that serve as strong lipid-binding domains. It also has several domains that could serve as binding sites for heparin-like molecules and form the basis for some of the cell surface interactions of the apo B-containing lipoproteins. In addition, apo B-100 contains an LDL receptor binding domain (amino acids 3100–3400), which allows the specific uptake of LDL by the LDL receptor.

Apolipoprotein B-48

The amino terminal 48% of apo B-100 forms apolipoprotein B-48 (apo B-48) (molecular weight 240 kDa). This apolipoprotein is produced from the *APOB* gene in the intestine by an mRNA editing process; a cytidine deaminase, *APOB* mRNA editing enzyme complex 1 (apoBec-1), binds to and acts on the cytosine molecule at base 6666 of the mRNA to form a uracil. The editing enzyme complex is only found in intestinal epithelial cells. Its action results in the glutamine 2153 triplet of CAA being converted into the stop codon, UAA. Thus protein synthesis is prematurely terminated at amino acid 2152 and, as a result, apo B-48 does not contain the LDL receptor binding domain present in apo B-100.

Apolipoprotein C

There are three apolipoprotein Cs, all of which are synthesized in the liver. In plasma, they transfer between the

triglyceride-rich lipoproteins (chylomicrons, VLDL and their remnants) and HDL.

Apolipoprotein C-I

Apolipoprotein C-I (apo C-I) (molecular weight 7kDa) forms a minor component of VLDL, IDL and HDL; it acts as an activator of LCAT.

Apolipoprotein C-II

Apolipoprotein C-II (apo C-II) (molecular weight 9kDa) is a component of chylomicrons and VLDL, in which it functions as an activator of LPL. Apo C-II is also found in IDL and HDL.

Apolipoprotein C-III

Apolipoprotein C-III (apo C-III) (molecular weight 9kDa) is synthesized mainly in the liver and, to a lesser extent, the intestine. It forms a major structural component of VLDL but is also present in chylomicrons and HDL. It acts as an inhibitor of LPL, and has more recently been shown to promote hepatic assembly and secretion of VLDL. Apo C-III also inhibits hepatic uptake of chylomicron and VLDL remnant particles, possibly by preventing interaction of apo E on these remnant particles with the hepatic receptor. High plasma apo C-III concentrations are associated with high plasma triglyceride concentrations.

Null mutations have been reported which are associated with low plasma triglyceride and LDL, and high HDL concentrations. However, the *APOC3* gene is close to the *APOA1* gene, and both are deficient in some forms of apo A1 deficiency, which cause low plasma HDL and triglyceride concentrations.

Apolipoprotein D

Apolipoprotein D (apo D) (molecular weight 33 kDa) is a lipoprotein-associated glycoprotein, forming a minor component of HDL, VLDL, IDL and LDL. It transports small hydrophobic ligands, including sterols and cholesterol. Apo D is associated with increased activity of lipoprotein lipase, and missense mutations cause elevated triglycerides. The apo D concentrations in the hippocampus and cerebrospinal fluid (CSF) of patients with Alzheimer disease are increased.

Apolipoprotein E

Apolipoprotein E (apo E) is a 299 amino acid glycoprotein (molecular weight 34kDa) synthesized by the liver and found in all classes of lipoproteins except LDL. It is involved in the control of chylomicron and VLDL remnant removal from the circulation. It also has antioxidant properties, and controls the efflux of cholesterol from cells, together with apo A-I. Apo E is a polymorphic protein: three common isoforms occur, which can be separated by isoelectric focusing and are designated apo E2, apo E3 and apo E4. Apo E2 differs from apo E3 by only a single amino acid, Cys being substituted for Arg

at residue 158. Apo E4 also differs from apo E3 by only a single amino acid, Arg being substituted for Cys at residue 112. The apo E3/E3 phenotype is the most common, comprising 50–70% of the population, whereas the apo E2/E2 phenotype is the least common, occurring in about 1% of the population.

The apo E3 isoform is associated with normal chylomicron and VLDL metabolism. The E2 isoform does not function as an effective ligand for the receptor-mediated uptake of remnant particles, having less than 2% of normal apo E3 binding to the LDL receptor. As a result, remnant lipoproteins tend to accumulate in the plasma of individuals homozygous for apo E2/E2. The E4 isoform is associated with higher concentrations of LDL-cholesterol than E3. Apo E4 homozygotes are also at increased risk of Alzheimer disease. Apo E synthesis is upregulated to aid cell repair in response to cellular stress or injury. Apo E4 is more susceptible to proteolytic cleavage than E2 or E3. This results in accumulation of intracellular fragments that cause changes in the cytoskeleton and the formation of neurofibrillary tangles.

Apolipoprotein M

Apolipoprotein M (apo M) (molecular weight 26kDa) fulfills the criteria for being an apolipoprotein as it is not found free in plasma but is predominantly associated with HDL. Like apo D, it is a member of the lipocalin family of proteins, which contain a binding domain for small lipophilic ligands, and may therefore have a role in transport of small lipid molecules. Although apo M is found in association with only 5% of HDL particles, where it may potentiate the antioxidant effect of HDL, its concentration is positively correlated with cholesterol concentration, suggesting that it may also have a role in cholesterol metabolism.

Apolipoprotein(a)

This apolipoprotein (apo(a)) is a large glycosylated protein of variable size (molecular weight 200–800 kDa). It contains multiple kinks in the polypeptide chain that are termed 'kringles'. Apo(a) is a homologue of plasminogen; it contains a single copy of plasminogen kringle 5, multiple copies of plasminogen kringle 4 and an inactive protease domain. Kringle 4 shows wide variation in the number of repeats within the apo(a) molecule. In a subpopulation of LDL particles, apo(a) forms disulphide bridges with apo B-100 to form a distinct lipoprotein class termed lipoprotein(a) (Lp(a)) (see p. 708). The function of apo(a) is unknown. It has a strong homology with plasminogen and may interfere with fibrinolysis.

CHOLESTEROL ABSORPTION

Cholesterol of both dietary and biliary origin is absorbed in the upper jejunum facilitated by a specific transporter protein, the Niemann–Pick C1-like 1 (NPC1L1) protein. This has 50% homology with the product of the *NPC1* gene, the Niemann–Pick C1 disease protein, which is involved in intracellular cholesterol trafficking. Niemann–Pick C1-like 1 protein transports plant sterols as well as cholesterol.

In small intestinal enterocytes, dietary cholesterol is released by the action of lysosomes and selectively esterified by acyl-CoA cholesterol acyltransferase 2 (ACAT2) prior to incorporation into chylomicrons. Any excess free cholesterol in enterocytes, together with any absorbed plant sterol, is excreted back into the intestinal lumen. The apical excretion of cholesterol and other sterols back into the gut lumen is a function of two hemitransporters of the ATP binding cassette transporter family, *ABCG5* and *ABCG8*. These are also responsible for the excretion of cholesterol and other sterols into the bile by hepatocytes. Expression of *ABCG5* and *ABCG8* in hepatocytes and enterocytes is under the control of the liver X receptor (LXR) and is induced by cholesterol feeding.

Cholesterol absorption reflects the imbalance between the movement of cholesterol across the brush border into enterocytes and its excretion by enterocytes back into the intestinal lumen. In normal individuals, the rate of cholesterol absorption correlates with the plasma sterol concentrations. The rate of cholesterol absorption is a predictor of benefit of statin treatment and response to ezetimibe (a selective inhibitor of NPC1L1). Those with higher rates of cholesterol absorption show reduced CVD benefit from statins, and increased cholesterol lowering with ezetimibe.

β -Sitosterolaemia, also referred to as phytosterolaemia, is an autosomal recessive disorder caused by mutations in *ABCG5* or *ABCG8*, in which there is increased absorption of dietary non-cholesterol sterols, and consequently an increased concentration (30–100-fold) of these sterols in blood. A ‘normal’ Western diet contains similar amounts (200–500 mg/day) of cholesterol and of non-cholesterol sterols (mainly plant but also some fish sterols). Approximately 55% of the cholesterol is usually absorbed, but normally less than 1% of the non-cholesterol sterols. In the absence of a functional *ABCG5/8* heterodimer, subjects overabsorb cholesterol and non-cholesterol sterols from enterocytes, because these transporters usually excrete some cholesterol and almost all non-cholesterol sterols back into the gut lumen. They also fail to excrete cholesterol and non-cholesterol sterols into the bile. These sterols become deposited at various sites in the body causing tendon xanthomas and premature atherosclerosis. Clinical features also include arthritis and haemolytic episodes. Treatment comprises a diet low in dietary sterols. Ezetimibe is also effective by inhibiting sterol absorption via NPC1L1.

TRIGLYCERIDE DIGESTION

Fat (triglyceride) digestion starts in the stomach, where lingual and gastric lipases hydrolyse 25–30% of ingested triglycerides into diglycerides and free fatty acids. In the duodenum, partially digested lipids mix with bile and pancreatic secretions. The latter contains a mixture of enzymes, including carboxyl ester lipase and pancreatic triglyceride lipase, which are capable of further hydrolysing dietary lipids to monoglycerides, free fatty acids, glycerol and cholesterol. The bile acids serve to solubilize the lipids and result in the formation of micelles, the contents

of which are absorbed by enterocytes. The absorption of monoglycerides is by passive diffusion. Within the enterocytes, triglycerides are resynthesized from free fatty acids and either monoglycerides or glycerol.

BILE ACID METABOLISM

Approximately 500 mg of cholesterol is converted, in the liver, to bile acids each day. This replaces the bile acids lost in the stools and represents approximately 5% of the bile acid pool, as the enterohepatic circulation is 95% efficient. The synthesis and secretion of bile acids, together with hepatic cholesterol secretion into bile, represents the major pathway for the elimination of cholesterol from the body. The synthesis of the full complement of bile acids from cholesterol requires 17 enzymatic steps. It is under negative feedback control: accumulation of bile acids leads to reduced activity of the key enzymes 7 α -hydroxylase and sterol 12 α -hydroxylase. The products of the bile acid synthetic pathway are the primary bile acids, cholic and chenodeoxycholic acids. In the gut, these primary bile acids are converted into secondary and tertiary bile acids by the action of anaerobic bacteria. The bile acids present in bile are a mixture of primary, secondary and tertiary, the latter being a consequence of enterohepatic circulation.

A number of enzymatic defects in the synthetic pathway of bile acids have been described; in general, the earlier in the pathway, the earlier in life clinical problems become manifest and the greater their severity. One of these disorders, due to defective sterol 27-hydroxylase activity, is cerebrotendinous xanthomatosis (CTX). In this condition, cholesterol and its 5 α -reduced derivative, cholestanol, accumulate in the blood and tissues. As a result of this accumulation, affected individuals develop tendon xanthomas like those found in familial hypercholesterolaemia. In CTX, however, these sterols also accumulate in myelin sheaths, leading to progressive neurological dysfunction.

The bile acids act to solubilize fats during absorption from the gut, but they also have a solubilizing action in the bile. If the proportions of bile acids, cholesterol and phospholipid in the bile are disturbed, there is an increased risk of gallstones being formed.

LIPOPROTEIN METABOLISM

Lipoprotein metabolism is summarized in [Figure 37.8](#).

Assembly of apolipoprotein B-containing lipoproteins

The assembly of the apo B-containing lipoproteins requires the coordinated synthesis of apo B and lipids. First, apo B has lipid added to it by microsomal triglyceride transfer protein (MTP) as it is being synthesized in the endoplasmic reticulum, to form a nascent lipoprotein particle. Further triglyceride may be added, forming VLDL2. In the next step, VLDL2 is exported from the endoplasmic reticulum by a membrane associated protein

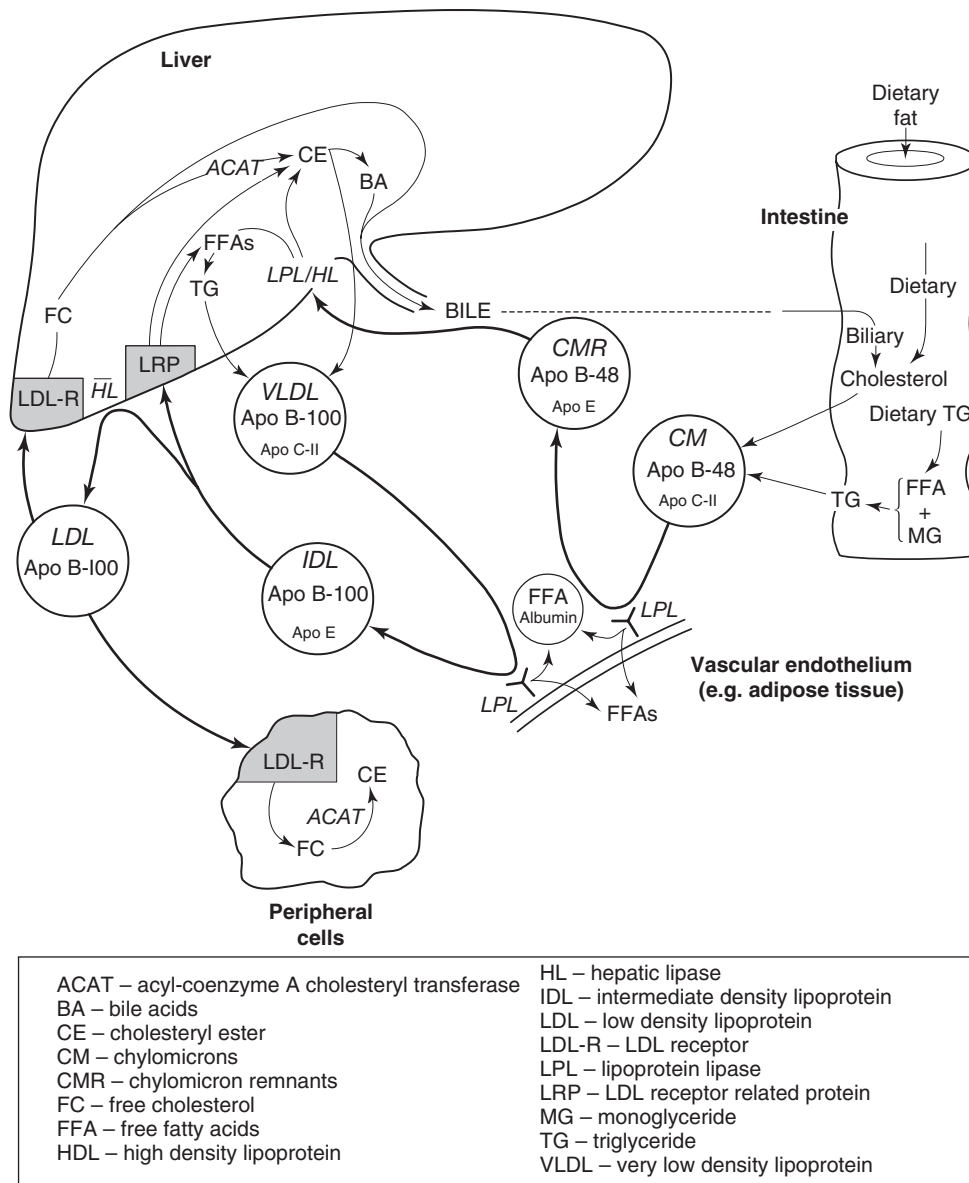


FIGURE 37.8 ■ Outline of the metabolism of the apo B-containing lipoproteins.

complex comprising a coatome protein (COPII) and a GTPase (sar1b). The resulting sar1/COPII vesicles then fuse with the Golgi apparatus. Apo A-IV, apo C-III and apo A-I are added to the surface of the lipoprotein. In the presence of large amounts of free fatty acids and triglycerides, phospholipase D1 and extracellular signal-regulated kinase 2 (ERK2) increase the formation of lipid droplets that deliver lipids to the Golgi apparatus and promote the further addition of lipid to VLDL2, which is thereby converted to VLDL1.

In healthy individuals, most VLDL particles are small and relatively triglyceride-poor VLDL2. In conditions such as insulin resistance and type 2 diabetes there is increased production of larger, triglyceride-rich VLDL1, which in turn generate atherogenic remnants, small dense LDL particles and triglyceride rich HDL particles that are more readily catabolized, leading to low concentrations of circulating HDL.

Apolipoprotein B is constitutively synthesized, but the rate of production of apo B-containing lipoproteins is related to the availability of fatty acids for triglyceride synthesis. When there is a plentiful supply of fatty acids, the majority of the apo B that is synthesized is incorporated into lipoproteins; however, when fatty acids are in short supply, the apo B is degraded intracellularly.

Once lipoprotein particles are fully formed, they are transferred to the cell surface for secretion. Very low density lipoprotein particles (200nm) and chylomicrons (1000nm) are large in the context of classic transport vesicles (range 50–80nm). The intracellular transport of chylomicrons and some VLDL particles relies on sar1/COPII vesicles. Defects in the *SARA2* gene that codes for sar1b result in chylomicron retention disease, in which there is an inability to secrete chylomicron particles. Apo B-48 is absent from the circulation, but apo B-100 containing lipoproteins are still present, although in reduced amounts,

as sar1b is also involved in later stages of the VLDL pathway. Newly synthesized VLDL is not necessarily secreted from the cell, and may undergo immediate degradation. Sortilins (see p. 721) are now recognized to be involved in routing newly synthesized lipoproteins either to intracellular degradation or for export into the circulation.

Exogenous pathway

Enterocytes absorb dietary cholesterol and triglyceride from the gut in the form of free cholesterol, fatty acids and monoglycerides. After re-esterification, cholesteryl esters and triglycerides (containing fatty acids of chain length C_{14} and greater) are incorporated into the cores of chylomicron particles. Enterocytes synthesize apo B-48 (the major protein of chylomicrons), apo A-I, apo A-II and apo A-IV that, together with phospholipids, form the surface layer of chylomicrons. Apolipoprotein B-48 is essential for chylomicron secretion. Secretory vesicles bud off from the Golgi apparatus and migrate to the basolateral regions of enterocytes. Here, they fuse with the plasma membrane and release chylomicrons into the intestinal lymphatics. The chylomicrons pass in the lymphatics to the thoracic duct and enter the circulation via the left subclavian vein.

There is only one apo B-48 molecule per chylomicron particle and it remains with the particle throughout its life until, as a chylomicron remnant, it is taken up by the liver. In contrast, multiple copies of the other apolipoproteins are present in a single chylomicron particle. These other apolipoproteins do not remain with a given chylomicron particle throughout its life but are exchanged with other lipoproteins. From the time of secretion, chylomicrons undergo constant modification, gaining apo C-II, apo C-III, apo E, phospholipids and cholesterol from HDL. The acquisition of apo C-II allows chylomicrons to interact with LPL, which is sited on the vascular endothelium, especially in adipose tissue and muscle. Lipoprotein lipase acts extracellularly to hydrolyse triglycerides within the chylomicron cores; the fatty acids thus released can be either utilized as an energy source or re-esterified and stored in adipose tissue as triglycerides. As hydrolysis proceeds, the cores of the chylomicrons reduce in size and excess surface components – phospholipids, free cholesterol, apo C-II and apo C-III – are transferred back to HDL. Apolipoprotein C-III may exert a modulating effect on LPL-catalysed chylomicron hydrolysis. Both apo C-II and LPL are necessary for normal chylomicron catabolism. Continuing loss of apo C-II as the mass of individual chylomicron particles is reduced eventually prevents further interaction with LPL, and chylomicron remnant particles are generated. These remnants have a relatively high content of cholesteryl esters and apo E.

Chylomicron remnants are normally cleared quickly from the plasma by the liver. They enter the space of Disse through the fenestrated sinusoidal endothelium, together with lipoprotein lipase. The space of Disse is rich in heparan sulfate proteoglycans (HSPRG) as well as hepatic lipase (HL) and apo E, which are secreted by hepatocytes. The proteoglycans and apo E bind the remnants that are further catabolized by HL and LPL, becoming enriched in apo E and further depleted in

triglycerides. The remnants may then be taken up in a process directly mediated by HSPRG or via apo E binding to the LDL receptor related protein (LRP). This LRP mediated uptake may be of remnant particles alone, or of remnants bound to HSPRG. The LDL receptor is also able to uptake chylomicron remnants via apo E binding in a process independent of HSPRG.

The cholesteryl esters delivered to the liver by chylomicron remnant particles may be utilized for the synthesis of bile acids or membranes or be secreted in VLDL, while the apo B-48 undergoes degradation.

In summary, chylomicron metabolism essentially comprises two steps. In the first step, most of the fatty acids derived by peripheral lipolysis of chylomicrons enter adipocytes for storage as triglycerides or other cells for oxidation for energy production. A small fraction of the released fatty acids is bound to plasma albumin and transported in the blood to the liver and other tissues. The second step involves remnant particles delivering the remaining triglycerides and almost all the cholesterol to the liver.

Lipolysis in adipose tissue

Triglycerides in white adipose tissue, far from being an inert store, are continuously undergoing lipolysis and re-esterification. In the fasting state, and at times of increased energy demand, free fatty acids are released into the circulation and transported to other tissues. The lipolysis of triglycerides in white adipose tissue is initiated by adipose triglyceride lipase (ATGL). Its action results in the production of diglycerides and free fatty acids, but it has very little hydrolytic activity towards diglycerides. Hormone sensitive lipase, which until recently was thought to be the rate-limiting enzyme involved in the release of free fatty acids from adipose tissue, is actually rate limiting for the hydrolysis of diglycerides rather than triglycerides. The monoglycerides resulting from the action of HSL are acted on by a third enzyme, monoglyceride lipase (MGL).

Adipocyte lipolysis is controlled by a number of lipolytic and antilipolytic hormones, including catecholamines and insulin.

Endogenous pathway

Hepatocytes are the originators and often also the acceptors of particles involved in the endogenous pathway, which shows many points of similarity with the exogenous pathway.

The liver secretes VLDL, a triglyceride-rich lipoprotein. The triglycerides are produced either de novo by hepatocytes or taken up from the plasma. These triglycerides, together with cholesterol derived from chylomicron remnants or from de novo synthesis, are secreted with phospholipids and apo B-100 as nascent VLDL. Some apo C-I, apo C-II, apo C-III and apo E are also present in the nascent VLDL particle, but the majority of these apolipoproteins are probably acquired from HDL within the circulation in the same way as in the exogenous pathway. In situations where there is an excess of hepatic triglyceride, large VLDL particles are secreted. Large VLDL particles are also secreted in familial hypertriglyceridaemia, whereas

in familial combined hyperlipidaemia (see p. 724), the rate of VLDL secretion is increased but a relative scarcity of triglycerides ensures that the individual VLDL particles are smaller and relatively poor in triglycerides.

The initial metabolic transformation of VLDL is a progressive LPL-mediated lipolysis analogous to the process involving chylomicrons. This requires apo C-II and produces cholesteryl esters and apo E-rich remnant particles. As in chylomicron metabolism, the surface components of VLDL are transferred back to HDL as the cores shrink. VLDL remnants comprise small VLDL particles and IDL. About half are cleared by the liver in a process that involves uptake by LRP, which recognizes apo E. The remainder undergoes further hydrolysis by HL to form LDL.

Low density lipoprotein is the major cholesterol-carrying lipoprotein in the plasma and usually accounts for 70% or more of the total plasma cholesterol. Virtually the only protein contained in the LDL particle is a single molecule of apo B-100, which acts as the ligand for the LDL receptor. LDL receptors are present on hepatocytes as well as the cells of peripheral tissues. Approximately 50% of plasma LDL uptake by the LDL receptor-mediated mechanism is hepatic. The major determinant of plasma LDL-C concentration is the number of functional LDL receptors.

Low density lipoprotein receptors recognize both apo B-100 on LDL and apo E on remnant particles and HDL. Once a lipoprotein has been bound to the receptor, the receptor-lipoprotein complex localizes in the coated pit region from where it is internalized by endocytosis. The LDL receptor is recycled while the lipoprotein undergoes lysosomal degradation to unesterified cholesterol

and amino acids. The cholesterol thereby released is available for further metabolic transformations as well as to regulate the transcription and/or translation of the HMG-CoA reductase and the LDL receptor genes. The cholesterol may be re-esterified by the action of ACAT and stored, or may be utilized for bile acid, steroid or membrane synthesis.

Hepatic cholesterol trafficking

The liver is the key organ for the regulation of cholesterol; not only is it responsible for the majority of cholesterol synthesis but it also acquires cholesterol from all lipoprotein classes. The cholesterol secreted into bile is mostly derived from lipoproteins with only a small contribution coming from de novo hepatic synthesis or hepatic cholesteryl ester stores. This involves preferential trafficking of lipoprotein-derived cholesterol, which involves multiple cholesterol transport-related gene products, the expression of which is regulated by the concerted activity of sterol-activated transcription factors.

High density lipoprotein metabolism (see Fig. 37.9)

Assembly of lipoproteins

Apo A-I is exported into the circulation by liver and intestinal cells. Here it may combine with phospholipids to form nascent (discoidal or pre- β_1) HDL.

Pre- β_1 HDL may also be formed from spare surface components of triglyceride rich lipoproteins (VLDL, chylomicrons, and their remnants) following the action of lipases.

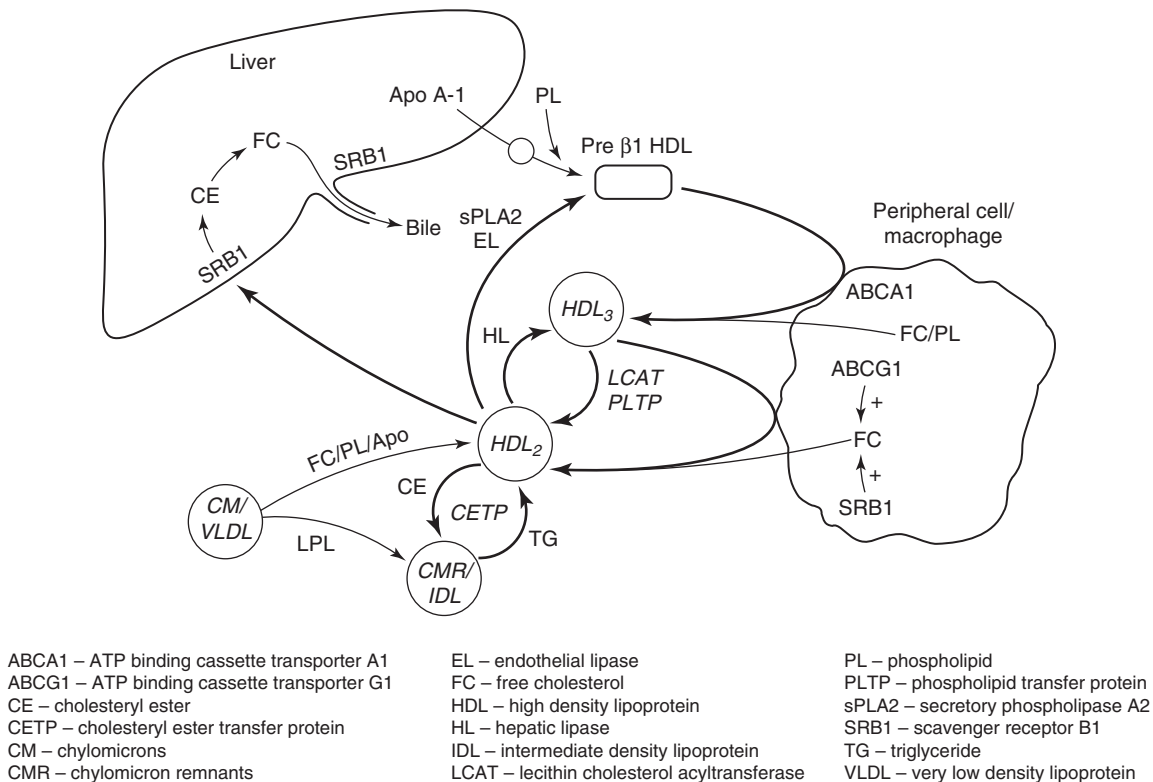


FIGURE 37.9 ■ Outline of HDL metabolism and its role in reverse cholesterol transport.

Cholesterol efflux

Except in steroidogenic tissues (adrenals and gonads), the catabolism of cholesterol in peripheral cells is limited to its partial degradation by 27 α -hydroxylase. Excess free cholesterol has to be removed. This occurs by an active process, or by diffusion. Free cholesterol is actively exported by ATP binding cassette transporter A1 (ABCA1), which preferentially binds pre- β_1 HDL. Expression of ABCA1 is regulated by the liver X receptor/retinoid X receptor (LXR/RXR) system; the physiological ligand of LXR is oxysterol, the cellular concentration of which increases in parallel with that of cholesterol.

Diffusion is non-specific efflux of cholesterol from cell surface membranes to an extracellular 'acceptor'. All HDL particles can act as acceptors of free cholesterol. This is a bidirectional process. Net efflux is increased by the activity of LCAT, which esterifies free cholesterol in lipoproteins, thereby increasing the concentration gradient between HDL and the cell membrane. Cholesterol efflux by diffusion is not as dependent on the activity of LCAT as previously thought, as it may also be facilitated by the SRB1 and the ATP binding cassette transporter G1 (ABCG1). The SRB1 binds preferentially to larger (HDL2) particles. In peripheral cells, SRB1 forms a hydrophobic channel through which free cholesterol can readily diffuse, enhancing the rate of efflux. The ABCG1 increases the concentration of free cholesterol within the plasma membrane and makes it more readily desorbed. This promotes the step which is normally rate limiting in cholesterol efflux by diffusion. Cholesterol efflux capacity correlates inversely with atherosclerosis, after adjustment for plasma HDL concentration. However, although both plasma apo A-I and HDL predict efflux capacity, they account for <50% of the variation seen. This emphasizes the importance of overall flux through the pathway of reverse cholesterol transport rather than simply plasma HDL concentration in the development of atheroma.

Reverse cholesterol transport

Free cholesterol exported from peripheral cells or macrophages combines with apo A-I to form 'nascent' pre- β_1 HDL. Esterification of free cholesterol on the surface of the HDL particle by LCAT produces hydrophobic cholesteryl esters that migrate to the centre of the particle, releasing capacity to accept more free cholesterol, and transforming it into a larger, more spherical particle (HDL3). HDL3 accepts more cholesterol by active export or facilitated diffusion, and LCAT further increases particle size. More free cholesterol, phospholipids and apolipoproteins (C-II, C-III, E) are also transferred from triglyceride-rich lipoproteins (TGRLs) and their remnants as triglycerides are removed from these particles by the action of LPL. The HDL particle increases in size to become HDL2, which contains a greater proportion of apo A-II and is the preferred ligand for the hepatic SRB1 receptor. When HDL2 is bound, cholesteryl esters are taken up into the cell, converted to free cholesterol and

may then be excreted in bile. Scavenger receptor B1 is present on the basolateral membrane of hepatocytes where it promotes influx of cholesterol, and on the canalicular membrane where it promotes efflux to bile. Cholesterol taken up into the hepatocyte by SRB1 may also be trafficked across the cell and pumped into bile by ABCG5/G8 at the canalicular membrane.

High density lipoprotein 2 may also transfer cholesterol to TGRLs in exchange for triglycerides, catalysed by CETP. This potentially 'diverts' cholesterol from the 'normal' reverse cholesterol transport pathway into potentially atherogenic lipoproteins, and may therefore be proatherogenic. However, in transferring cholesterol to the VLDL/LDL pathway, it allows for cholesterol clearance via hepatic LDL receptors, and may theoretically be antiatherogenic in situations where clearance of cholesterol from HDL via SRB1 is saturated. The triglycerides that enrich HDL as a result of this exchange are hydrolysed by lipases and this results in small dense HDL, which is more readily catabolized by the kidney, depleting plasma HDL. This CETP-mediated exchange is increased in the presence of increased circulating amounts of TGRLs and this accounts for the inverse relationship between plasma triglycerides and HDL. HDL2 is a substrate for hepatic and endothelial lipases (ELs) and secretory phospholipase A2 (sPLA2). These may regenerate pre- β_1 HDL or HDL3, but if HDL is triglyceride-enriched by excessive CETP activity then the HDL particle may disintegrate, promoting catabolism in the kidneys.

High density lipoprotein therefore exists as a complex interchange of lipoproteins. ABCA1, ABCG1, LCAT and LPL all increase particle size, while CETP, SRB1, HL, EL and sPLA2 tend to reduce size, with net transport of free cholesterol from peripheral cells to the liver. These interactions also have the potential to 'divert' cholesterol to atherogenic TGRLs, with loss of apo A-I, especially in conditions where TGRLs are increased.

High density lipoprotein cholesterol (HDL-C) shows an inverse relationship with atherosclerosis in epidemiological studies, although this is not a straightforward as previously thought. Very high concentrations of HDL-C are associated with increased risk of atherosclerosis, which could be due to HDL being 'trapped' as a 'dysfunctional' pool in plasma rather than reflecting an increased flux through the reverse cholesterol transport pathway. This highlights the importance of qualitative changes in HDL that may affect its function, and the overall net flux through the reverse cholesterol transport pathway (see p. 714) as well as total plasma concentration.

High density lipoprotein has direct anti-atherogenic properties including:

- downregulation of the expression of adhesion molecules on the surface of the vascular endothelium
- inhibition of platelet aggregation
- prevention of the inhibition of nitric oxide synthase by oxidized LDL
- upregulation of ATP binding cassette transporter A1 (ABCA1) and activation of intracellular signalling leading to stimulation of cholesterol efflux.

In addition, paraoxonase, an ester hydrolase that is transported bound to HDL, has antioxidant properties.

ENZYMES INVOLVED IN LIPOPROTEIN METABOLISM

Lecithin cholesterol acyltransferase

Lecithin cholesterol acyltransferase (LCAT) is a glycoprotein that has both a phospholipase A2 and an acyltransferase action. It is essential for the normal maturation, interconversion and rearrangements of all lipoprotein classes and is involved in reverse cholesterol transport. It is synthesized by the liver and circulates in the plasma reversibly bound to HDL and LDL. Lecithin cholesterol acyltransferase catalyses the esterification of free cholesterol on the surfaces of lipoproteins. The preferred substrate for LCAT is HDL, which contains apo A-I, its most potent activator. The ability of apo A-I to activate LCAT depends on the lipid composition of HDL, being most effective in small HDL particles, which have a low ratio of phospholipid to apo A-I. The cholesteryl esters formed by the action of LCAT expand the cores of the HDL particles. Lecithin cholesterol acyltransferase activity may be a rate limiting step in reverse cholesterol transport, making it a potential therapeutic target in the prevention of atherosclerosis, although evidence as to the effect of LCAT on the development of atherosclerosis is conflicting. Both recombinant LCAT and small molecule activators of LCAT activity are being explored as possible therapies.

Mutations in the *LCAT* gene cause either familial LCAT deficiency, when there is no LCAT activity or fish eye disease, with partial LCAT deficiency (see p. 729).

Lipases

Lipoprotein lipase

Lipoprotein lipase (LPL) is an extracellular enzyme that is bound by the glycosaminoglycan heparan sulphate to capillary endothelial cells. It is present in large amounts in the capillaries of adipose tissue and muscle, both skeletal and cardiac. Lipoprotein lipase belongs to the triglyceride lipase gene family, which also includes HL, EL, and pancreatic triglyceride lipase (PTL). It is the major lipolytic enzyme involved in the intravascular metabolism of the triglyceride-rich lipoproteins. Myocytes and adipocytes secrete LPL in a catalytically inactive form, which is then transported to the capillary endothelial surface.

Along with HL and EL, LPL is a homodimer. Interaction of the dimer with heparan sulphate on the endothelial surface serves to anchor and stabilize the LPL. LPL monomers that are catalytically inactive are found in the circulation in association with remnant particles and may play a role in enhancing their clearance. Presence of apo C-II is required for full activity.

Lipoprotein lipase catalyses the partial hydrolysis of the core triglycerides of chylomicrons and VLDL to monoglycerides and fatty acids. The fatty acids are taken up by the tissue and either re-esterified and stored (in adipose tissue), utilized as an energy source (in muscle) or secreted (in lactating breast tissue). The monoglycerides are further hydrolysed to glycerol and fatty acids.

Lipoprotein lipase binds to heparin, which results in its release into the circulation. This is used in the assay

of LPL activity (post-heparin lipolytic activity, PHLA). Lipoprotein lipase regulates the plasma concentrations of triglycerides and HDL. Individuals with low PHLA, such as those who are heterozygous for LPL deficiency, have high triglyceride and low HDL concentrations in plasma, and an increased risk of atherosclerosis.

A large number of mutations in the *LPL* gene have been described. Some are a cause of the familial chylomicronaemia syndrome (see p. 725), while others have less severe effects. It is estimated that 20% of patients with hypertriglyceridaemia are carriers of *LPL* gene mutations. The Asn291Ser mutation is present in 2–5% of Caucasians and is associated with a 31% increase in plasma triglyceride concentrations and an increased risk of coronary heart disease and type 2 diabetes. Asn291 is located in part of the molecule involved in homodimer formation, so it is likely that this mutation causes an increase in the relative amount of LPL present as inactive monomer. The risk of Alzheimer disease is also increased in Asn291Ser carriers.

Small amounts of LPL have been demonstrated to be present on arterial endothelium and also within the intima of arteries. LDL binds to LPL with an affinity similar to chylomicrons and VLDL. Although LDL is not the physiological substrate for LPL, the LPL present in the arterial wall may, by binding to LDL, increase the residence time of LDL in the arterial wall, thus promoting atherogenesis.

Hepatic lipase

This enzyme is a product of the same gene family as LPL. It is synthesized in the liver, where 95% remains, bound to the exterior of the hepatocyte membrane in the spaces of Disse; the remainder is found in the adrenals, ovaries and macrophages. HL has a role in the remodelling of chylomicron remnants, IDL, LDL and HDL and is involved in reverse cholesterol transport. Its concentration is inversely related to that of HDL. Its predominant activity on HDL is as a phospholipase, whereas in other lipoproteins it acts as a triglyceride lipase. In addition, HL may influence lipoprotein metabolism by means of its capacity to form a 'bridge' between lipoproteins and lipoprotein receptors.

Enzyme activity is lower in females than males and is modulated by insulin resistance, dietary fat intake, visceral obesity, physical activity, smoking and certain drugs. The size and buoyancy of LDL are inversely related to HL activity; when its activity is high, small, dense LDL particles are produced.

Like LPL, HL can bind to heparin, which results in its release into the circulation. This activity is utilized in HL assays.

Endothelial lipase

Endothelial lipase is the third member of the triglyceride lipase gene family and like the others it exists as a homodimer. It is produced by hepatic endothelial cells, and unlike HL and LPL it functions at the site of synthesis. It is also found in macrophages. Its predominant activity is as a phospholipase A1. Transcription is induced by

cytokines and physical forces, which has led to the suggestion that it may play a role in the development of atheroma. Overexpression is associated with reduced plasma concentrations of HDL.

Lipase maturation factor 1

This membrane bound protein located in the endoplasmic reticulum is essential to the folding and assembly of the dimeric lipases (LPL, HL, EL). Loss of function mutations in *LMF1* cause failure in the assembly of homodimers, thus preventing these lipases from attaining normal catalytic activity. This produces combined lipase deficiency with marked hypertriglyceridaemia.

Pancreatic triglyceride lipase

This enzyme, synthesized by pancreatic acinar cells, is essential for the digestion of dietary triglycerides. Enzyme activity requires formation of a complex with its protein cofactor, colipase. Following emulsification, the second step in the digestion of dietary triglycerides involves their hydrolysis to fatty acids and monoglycerides, which are incorporated into micelles with bile salts. Pancreatic triglyceride lipase (PTL) accounts for the majority of triglyceride lipase activity in the upper small intestine of adults.

The pancreatic acinar cells also synthesize two pancreatic lipase-related proteins (PLRP1, PLRP2) that have a high degree of structural and sequence homology to PTL. PLRP2 may play a role in dietary fat digestion in the newborn but no catalytic activity has, as yet, been demonstrated for PLRP1.

Hormone sensitive lipase

This enzyme is not related to any of the above lipases. It is a neutral intracellular lipase that has catalytic activity against diglycerides, monoglycerides and cholesteryl esters. It is predominantly located in adipose tissue. It is also expressed in the adrenals, testes and ovaries and, to a lesser extent, in skeletal muscle, cardiac muscle, macrophages and pancreatic islets. The activity of HSL is rate limiting in the release of fatty acids from adipose tissue. It is subject to a complex control mechanism, which involves both prolipolytic (ACTH and catecholamines) and antilipolytic (insulin) hormones, via cyclic AMP-dependent protein kinase phosphorylation. HSL interacts with adipocyte lipid binding protein (ALBP), which is a member of a family of lipid binding proteins that bind fatty acids and other hydrophobic ligands. In vitro, monoglycerides and fatty acids exert feedback inhibition on HSL; the interaction with ALBP, which sequesters these products, may prevent this happening in vivo. HSL is also responsible for most cholesteryl ester hydrolase activity in the adrenals.

Carboxyl ester lipase

This enzyme is a non-specific lipase capable of hydrolysing cholesteryl esters, triglycerides, diglycerides, monoglycerides, phospholipids, lysophospholipids and

ceramide. It is the only enzyme in the gut with cholesteryl esterase action; all its other actions can also be effected by other enzymes, for example the triglyceride lipase action of pancreatic triglyceride lipase. Most carboxyl ester lipase (CEL) is synthesized by pancreatic acinar cells but it is also synthesized by the liver, macrophages and lactating breast tissue. In the intestine, CEL is attached to the brush borders of enterocytes by means of its heparin binding domain. It is taken up by enterocytes, possibly by endocytosis and may be involved in the intracellular processing of lipids during the assembly of chylomicrons. The fact that CEL acts on ceramide is important because ceramide can disrupt intracellular lipid and protein trafficking, resulting in a block in the assembly of large lipoproteins and their secretion by enterocytes; CEL prevents ceramide disrupting this process.

Vitamin A is ingested in the form of retinyl palmitate or retinyl acetate, which need to be hydrolysed prior to absorption. It is likely that this hydrolysis is a function of CEL in the neonatal period, although in the adult, other enzymes may be involved.

Lysosomal acid lipase

Lysosomal acid lipase (LAL) is secreted as a 399-amino acid precursor, which includes a 27-amino acid signal peptide for transport across the membrane of the endoplasmic reticulum. Further processing, including N-glycosylation in the endoplasmic reticulum and the attachment of mannose 6-phosphate residues in the Golgi, leads to lysosomal targeting. LAL is responsible for the breakdown of cholesteryl esters and triglycerides that are delivered to lysosomes as a result of receptor-mediated uptake of lipoproteins.

Wolman disease is an autosomal recessive disease that presents in the neonate and is characterized by hepatosplenomegaly, steatorrhoea and abdominal distention, results from complete lack of LAL activity, causing massive accumulation of both cholesteryl esters and triglycerides in macrophages throughout the body. Affected children usually die before their first birthday. Enzyme replacement therapy has now been developed but its efficacy remains to be fully established.

Cholesteryl ester storage disease is caused by partial deficiency of LAL. It is a rare autosomal recessive disorder, characterized by hepatomegaly, abnormal liver function tests, hypercholesterolaemia and premature atherosclerosis. The hepatomegaly is a consequence of accumulation of cholesteryl esters and triglycerides in both hepatocytes and Kupffer cells.

Phospholipase A2

Phospholipase A2 (PLA2) is one of a family of phospholipases, enzymes that hydrolyse phospholipids. It exists in five forms. Three are found within cells. Two are secreted, one of which, lipoprotein-associated PLA2 (LpPLA2) is also known as platelet activating factor acetyl hydrolase (PAF-AH). Lipoprotein-associated PLA2 promotes the hydrolysis of oxidized phospholipids in lipoproteins,

generating lysophospholipids and proinflammatory oxidized fatty acids.

There is a correlation between plasma LpPLA₂ concentration and atherosclerosis. It is unclear whether this association is causal, but inhibition of LpPLA₂ is of potential therapeutic value, and clinical trials are underway.

Acyl-CoA:cholesterol acyltransferase

Acyl-CoA:cholesterol acyltransferase (ACAT) catalyses the esterification of cholesterol. Two forms have been identified: ACAT1 is found ubiquitously in the endoplasmic reticulum of cells; ACAT2 is predominantly found in the endoplasmic reticulum of the liver and intestine. ACAT1 is sensitive to the degree of membrane cholesterol enrichment, and it is thought that its function is to maintain the cholesterol content of cell membranes at an optimal level by catalysing the esterification of excess free cholesterol. ACAT2 is responsible for the secretion of cholesteryl esters into the apo B-containing lipoproteins.

TRANSFER PROTEINS INVOLVED IN LIPOPROTEIN METABOLISM

Cholesteryl ester transfer protein (CETP)

This is a hydrophobic glycoprotein that is secreted mostly by the liver. In the blood, it is mainly bound to HDL. It facilitates the transfer of cholesteryl esters from HDL to the triglyceride-rich lipoproteins and LDL, and of triglyceride from triglyceride-rich lipoproteins to HDL. The extent of this exchange depends on the plasma concentration of triglycerides; the more triglyceride is transferred into HDL, the more is available as a substrate for HDL, and the greater the capacity for HDL to become more delipidated (and so more rapidly degraded). This is the explanation for the well-recognized inverse relationship between plasma triglyceride and HDL-cholesterol concentrations.

Cholesteryl ester transfer protein has been viewed as both anti- and pro-atherogenic. Its anti-atherogenic action is related to its role in reverse cholesterol transport: its pro-atherogenic action is related to its capacity to transfer cholesteryl esters from HDL to the atherogenic lipoproteins (VLDL, IDL and LDL). Under normal conditions, CETP-mediated cholesteryl ester transfer from HDL is predominantly to LDL. When the plasma concentration of VLDL is increased, as in type 2 diabetes, CETP-mediated cholesteryl ester transfer is preferentially directed towards the large VLDL particles, which consequently become cholesterol enriched and more atherogenic.

Although there is uncertainty about the potential effects of CETP activity on atheroma formation, it is a target for lipid-modifying drug therapy. Drugs that inhibit CETP may raise plasma HDL concentrations by over 100%. The results of clinical trials to see if this translates into effective reduction of clinical cardiovascular endpoints are awaited.

Phospholipid transfer protein (PTP)

This glycoprotein is a member of the same gene family as CETP. Along with other genes involved in lipid metabolism

the *PLTP* gene is under control of the liver X receptor (LXR). Phospholipid transfer protein is important in the remodelling of HDL, the generation of pre- β_1 HDL, the transfer of surface lipids from the triglyceride-rich lipoproteins to HDL during intravascular lipolysis and the facilitation of HDL-mediated efflux of phospholipids and cholesterol from cells. PTP activity is higher in hypertriglyceridaemic than normotriglyceridaemic individuals, and is correlated with the degree of insulin resistance. Its activity has also been found to be increased in subjects with familial combined hyperlipidaemia.

Fatty acid transport proteins

Fatty acid transfer proteins (FATPs) are membrane proteins that facilitate the uptake of fatty acids by cells. They also have acyl-CoA synthase activity. They form a family of six proteins that differ in tissue expression, intracellular localization and responsiveness to insulin. FATP1 and FATP4 have been implicated in insulin resistance. The former is highly expressed in adipose tissue and skeletal muscle whereas the latter is principally expressed in the small intestine. FATP2 is located in peroxisomes of the kidney and liver. It is important for the formation and oxidation of very long chain fatty acyl-CoA and is involved in the activation of bile acids in the liver; it is therefore sometimes referred to as very long chain acyl-CoA synthase. FATP3 is located in mitochondrial membranes. FATP5 is located in the membrane of the endoplasmic reticulum and is also involved in the activation of bile acids. FATP6 is located in the plasma membrane of cardiac myocytes, where it probably plays an important role in the supply of fatty acids as an energy source.

RECEPTORS INVOLVED IN LIPOPROTEIN METABOLISM

The LDL receptor

This receptor is a transmembrane protein that can be expressed by most cell types under certain conditions. The LDL receptor (LDLR) binds two apolipoprotein ligands, apo B-100 and apo E and, for this reason, is sometimes referred to as the B-100/E receptor. Uptake of LDL via the LDLR is mediated through its interaction with apo B-100. Although VLDL and IDL also possess apo B-100, it is not accessible for binding: IDL binds to LDLR via apo E. Subclasses of HDL containing apo E can also bind to the LDLR. Indeed, lipoproteins that contain multiple copies of apo E bind to the LDLR with much greater affinity than does LDL.

The *LDLR* gene is located on chromosome 19. The LDLR protein contains 843 amino acids and comprises five domains (see Fig. 37.10). At the amino terminal end is the ligand binding domain (exons 2–6), which mediates binding to apo B-100 or apo E. Next, there is a domain (exons 7–14) homologous to the epidermal growth factor (EGF) precursor. This contains three EGF repeats separated by a region of beta sheet folded in a specific orientation called a 'beta propeller'. At the acid pH found in

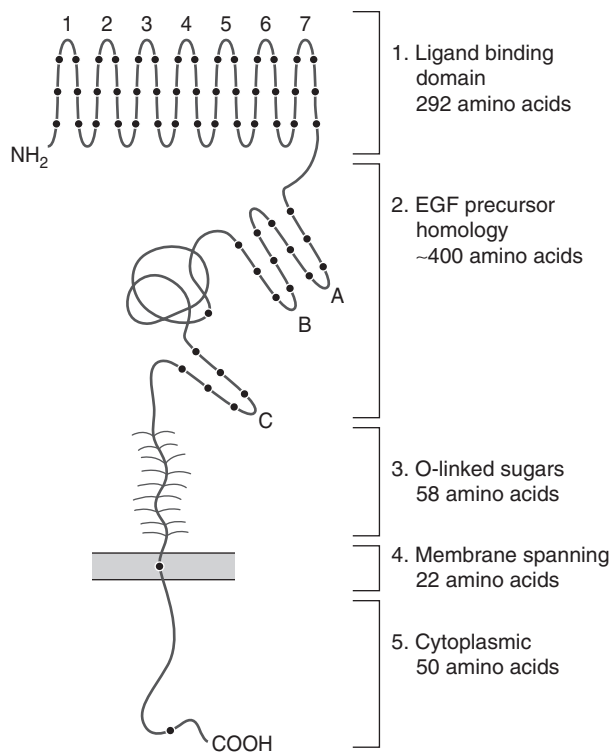


FIGURE 37.10 ■ The LDL receptor. EGF, epidermal growth factor.

endosomes, this domain undergoes a conformational change bringing the beta propeller adjacent to repeats 4 and 5 in the ligand binding domain and releasing bound LDL. It also serves to maintain the correct orientation of the ligand binding domain on cell surfaces. The specific orientation determined by this EGF domain is required for the binding of LDL to the receptor but not for the binding of apo E-containing lipoproteins. The third domain (exon 15) consists of a peptide chain enriched in serine and threonine to which sugar chains are attached. The fourth domain (exon 16) comprises hydrophobic amino acids that span the membrane and thus function as an anchor. The carboxy terminal portion constitutes the intracellular cytoplasmic domain (exons 17–18) and appears to be important in targeting the receptor to the coated pit region.

Expression of *LDLR* is under feedback control: in the steady state, the number of LDL receptors is controlled to allow uptake of sufficient cholesterol for cell growth and to balance losses. This feedback is by membrane bound transcription factors known as sterol regulatory element binding proteins (SREBPs) (see p. 721) and their associated insulin signalling (Insig) and SREBP cleavage activating proteins (SCAPs). The function of LDLRs is dependent not only on their binding to lipoprotein particles but also on their clustering in clathrin-coated pits for the subsequent internalization of the bound lipoproteins by endocytosis.

Low density lipoprotein receptor-associated protein 1 (LDLRAP1) binds the LDLR, clathrin and the clathrin-associated adaptin protein complex 2 (AP-2), and is essential for the endocytosis of LDLRs into hepatocytes. Mutations in the *LDLRAP1* gene cause the rare clinical phenotype of autosomal recessive familial hypercholesterolaemia.

LDL receptor-related protein

LDL receptor-related protein (LRP) is expressed by hepatocytes and has multiple ligands including apo E, HL and lipoprotein lipase. Following synthesis, LRP becomes associated with another protein, LDL receptor-related protein associated protein (LRPAP1), which acts as a chaperone to facilitate the passage of LRP to the cell surface. LRP is involved in the apo E-mediated uptake of remnant particles by the liver, and is sometimes referred to as the 'apo E receptor'.

The apo E4 polymorphism has an impact on the uptake of cholesterol to neuronal tissue by LRP. This may partially explain the increased risk of Alzheimer disease in those with the apo $\epsilon 4/\epsilon 4$ genotype.

Scavenger receptor class B type 1

This is a 509-amino acid protein comprising short N- and C-terminal cytoplasmic and transmembrane domains and a large extracellular loop. Its amino acid sequence is 30% homologous with that of CD36 (see below). Like CD36, scavenger receptor class B type 1 (SRB1) can bind to a number of lipoproteins and modified lipoproteins (HDL, LDL, VLDL, oxidized LDL). It is present in a number of tissues but is most highly expressed in tissues that process HDL cholesterol and cholesteryl esters for excretion or as metabolic substrates, i.e. the liver and steroidogenic tissues. SRB1 binds HDL and mediates the selective uptake of cholesteryl esters from the cores of HDL particles to the plasma membrane without entailing the uptake and degradation of the whole HDL particle. This process is dependent on apo A-I. These cholesteryl esters are then hydrolysed extralysosomally by tissue-specific neutral cholesteryl hydrolases; in the adrenals this may be by HSL.

Scavenger receptor class B type 1 is also able to mediate the efflux of unesterified cholesterol from cells to lipoproteins. Thus, SRB1 is involved at both ends of reverse cholesterol transport; initially with the efflux of free cholesterol from peripheral cells to HDL and at the end in the delivery of cholesteryl esters and unesterified cholesterol to the liver for secretion into bile. The membrane-containing protein PDZK1 (the PDZ group of proteins are named from the initial letters of the first three proteins found to share this structural domain), controls the basal hepatocyte and enterocyte SRB1 content, whilst in the adrenals, ACTH binding to the melanocortin 2 receptor stimulates the expression of SRB1. This in turn promotes the uptake of cholesteryl esters from HDL, which provides a pool of free cholesterol within the adrenals for synthesis of glucocorticoids. In the liver, SRB1 is downregulated by fibric acid derivatives. Scavenger receptor class B type 1 has been shown to mediate the binding and endocytosis of lipopolysaccharides, suggesting that it may have a role in their cellular uptake and clearance in sepsis.

There is evidence that loss of function mutations in the *SCARB1* gene that codes for SRB1 are associated with increased concentrations of HDL but a paradoxical increase in cardiovascular disease.

Other scavenger receptors

CD36 (which stands for Cluster of Differentiation 36) is a receptor found in caveolae in the cell surface plasma

membrane. Like SRB1 it is a Class B scavenger receptor. CD36 can bind native and oxidized LDL. Together with the scavenger receptors A1 and A2 (SR-AI, SR-AII) it mediates the avid uptake of oxidized LDL by endothelial macrophages.

Peroxisome proliferator-activated receptor family

Nuclear receptors play a key role in glucose and lipid homeostasis and in the inflammatory response. They are activated by a variety of ligands, including fatty acids. Among these receptors is the peroxisome proliferator activated receptor (PPAR) family. When they bind to a ligand, PPARs form heterodimers with the nuclear retinoid X receptor (RXR), which bind to response elements in the promoter regions of target genes, thus promoting their transcription. By this means, PPARs control the expression of genes that maintain lipid and glucose homeostasis. This makes them a potential therapeutic target.

The PPAR family includes PPAR α , PPAR γ and PPAR β/δ . The fibrate family of lipid-lowering drugs are activators of PPAR α and the thiazolidinediones, which are insulin sensitizing agents, are ligands for PPAR γ . The intravascular lipolysis of lipoproteins is promoted by fibrates through increased LPL activity and reduced expression of the LPL inhibitor, apo C-III. Fibrates also induce the expression of apo A-V, which impacts on plasma triglyceride concentrations and of apo A-I and apo A-II, leading to increased production of HDL. Activation of PPAR γ induces the expression of genes for proteins that control adipocyte fatty acid metabolism, including LPL and fatty acid transport proteins, thus facilitating the hydrolysis of lipoprotein triglycerides and promoting the storage of the released fatty acids as adipocyte triglycerides.

Other nuclear receptors

As well as PPARs, other nuclear receptors such as the liver X receptors (LXRs) and farnesoid X-activated receptor (FXR) have been demonstrated to exert control over lipid metabolism. The LXR is a sterol responsive nuclear receptor that is activated in response to excess of cholesterol within the cell. When activated, it triggers transcription of the inducible degrader of LDLR (IDOL); this causes ubiquitination of the cytoplasmic domain of LDLR, which promotes its destruction within the cell. Activated LXR also activates ABCA1 and ABCG1, which promotes cholesterol efflux from the cell. The net effect is to reduce cellular cholesterol uptake and increase efflux from the cell, this restoring sterol homeostasis.

OTHER PROTEINS INVOLVED IN LIPOPROTEIN SYNTHESIS, TRANSPORT AND METABOLISM

Microsomal triglyceride transfer protein

Microsomal triglyceride transfer protein (MTP) is a member of a group of proteins that are able to transfer of lipids between membranes. MTP is an 894 amino acid (97 kDa)

protein that forms a heterodimer with protein disulfide isomerase (PDI); the two subunits are held together by non-covalent interactions. This protein is present in high concentrations in the lumen of the endoplasmic reticulum and plays an important role in the maturation of the endoplasmic reticulum and its secreted proteins. Protein disulfide isomerase does not demonstrate isomerase activity when combined with MTP, but is required for retention of MTP in the endoplasmic reticulum at the site of apo B translocation. The promoter region of MTP is upregulated by cholesterol and downregulated by insulin. Microsomal triglyceride transfer protein binds and shuttles individual lipids between membranes.

Microsomal triglyceride transfer protein activity is essential for the synthesis of chylomicrons and VLDL. After the assembly of chylomicrons and VLDL within the lumen of the endoplasmic reticulum of enterocytes and hepatocytes respectively, they are transported to the Golgi apparatus and then secreted. Loss of function mutations in MTP cause abetalipoproteinaemia (see p. 724)

Microsomal transfer protein also transfers lipid to antigen presenting molecules on the surface of natural killer T cells. These mediate a number of autoimmune disorders, making MTP a pharmacological target, with inhibition having the potential to achieve both lipid lowering and anti-inflammatory effects. Inhibitors of MTP, which effectively lower plasma LDL-cholesterol concentrations, are in clinical trials, although it remains to be seen whether their use will be limited because of their tendency to cause an increase in hepatic steatosis.

ATP binding cassette transporter family

The ATP binding cassette (ABC) transporter family comprises more than 50 different proteins. Within this family, ABCG1, ABCG5, ABCG8 and ABCA1 are all sterol induced. ABCG5 and ABCG8 form a heterodimer, which mediates the excretion of absorbed plant sterols and cholesterol from enterocytes into the lumen of the gut and from hepatocytes into the bile. ATP binding cassette transporters A1 and G1 facilitate the export of cholesterol from peripheral cells, where its combination with apo A-I and phospholipids to produce HDL is the first step in reverse cholesterol transport.

Proprotein convertase subtilisin kexin 9

Proprotein convertase subtilisin kexin 9 (PCSK9) is one of nine members of the subtilisin protease family. The *PCSK9* gene is mainly expressed in liver and is regulated by sterol regulatory element binding proteins (see p. 721). The PCSK9 protein is synthesized as a 72 kDa protein that undergoes self-cleavage to a 63 kDa protein, which is found in the Golgi. Proprotein convertase subtilisin kexin 9 regulates LDLR activity by accelerating the degradation of mature receptors. It is expressed on the cell surface but also circulates free in plasma.

Gain of function mutations in the *PCSK9* gene are a rare cause of autosomal dominant hypercholesterolaemia (familial hypercholesterolaemia). Loss of function

mutations are associated with reduced plasma LDL-cholesterol concentration and lower than average risk of CVD. This makes both gene inactivation and immunological blockade of PCSK9 action attractive targets for lipid lowering therapy. Therapeutic agents are in clinical trials, and have been shown to reduce plasma LDL-cholesterol concentrations by >50%.

Sterol regulatory element binding proteins

Sterol regulatory element binding proteins (SREBPs) are membrane-bound transcription factors that activate genes involved in cholesterol synthesis. They provide the means by which cellular cholesterol exerts negative feedback on cholesterol synthesis. There are three SREBPs: SREBP-2 activates genes involved in cholesterol synthesis, the other two (SREBP-1a and SREBP-1c) have more effect on the genes involved in fatty acid synthesis. When cholesterol is abundant in a cell, the SREBPs remain in the endoplasmic reticulum associated with an escort protein, SREBP cleavage activating protein (SCAP), and the endoplasmic reticulum retention protein, insulin signalling protein (Insig). When a cell is relatively cholesterol-depleted, SCAP dissociates from Insig and results in the release of SREBP-SCAP from the endoplasmic reticulum. To exert their action, SREBPs must undergo two-step proteolytic cleavage in the Golgi to release their N-terminal fragment. These fragments travel to the nucleus where they bind to the steroid response elements (SREs) in the promoter regions of a number of the genes encoding enzymes involved in cholesterol synthesis, including squalene synthase and HMG-CoA reductase, and the *LDLR* gene.

When intracellular cholesterol content is again adequate, the proteolytic release of the N-terminal fragment is blocked and the N-terminal fragments in the nucleus are degraded.

Both *SREBP* genes contain intronic sequences that code for a micro-RNA (miR-33). Micro-RNAs (miRNAs) are small non-coding RNAs approximately 22 nucleotides in length that bind to the three prime untranslated region (3'UTR) of mRNAs, causing their reduced translation or destruction. Micro-RNA-33 reduces translation of mRNA for *ABCA1*, *ABCG1* and genes involved in fatty acid oxidation. This effect is complementary to the primary effect of SREBP on cholesterol and fatty acid synthesis and serves to further increase the accumulation of cholesterol and triglycerides within cells.

In the liver, insulin stimulates SREBP-1c production, and in states of hyperinsulinaemia with insulin resistance this may contribute to the accumulation of fat in hepatocytes, and drive increased VLDL synthesis. It is also possible that reduced *ABCA1* expression caused by miR-33 may account in part for the low HDL seen in insulin resistant states.

These recent insights into control of lipid metabolism by miRNAs makes them potential therapeutic targets, and initial studies have shown that HDL can be increased by antisense oligonucleotides designed to block miR-33.

Sortilins

Sortilins are receptors that regulate intracellular transport by shunting proteins through secretory or endocytic pathways. They share in common a 700 amino acid extracellular domain which forms a ten bladed beta propeller and acts as a ligand binding site. This is designated the vacuolar protein sorting ten protein domain (VPS10P).

Genome wide association studies suggest that two of these receptors, sortilin and sorting protein-related receptor with type A repeats (SORLA), have a role in lipid metabolism and atherosclerosis, and may account for variability in cardiovascular risk. Reduced quantities of SORLA are associated with increased smooth muscle cell migration and lipid loading of macrophages in atheromatous plaques. This effect may be mediated by SORLA binding of LPL and apo A-V on the cell surface. Overexpression of the sortilin gene, *SORT1* is associated with lower plasma LDL-cholesterol concentrations, either indirectly by reducing VLDL output from the liver, or directly by acting as an alternative hepatic receptor for LDL.

Glycosylphosphatidylinositol-anchored HDL-binding protein 1

Glycosylphosphatidylinositol-anchored HDL-binding protein1 (GPIHBP1) is a 184 amino acid protein that is a member of the GPI anchored Ly6 (lymphocyte antigen 6) group of proteins. It was originally identified as a protein capable of binding HDL. It is now recognized to be essential to the function of lipoprotein lipase. This protein binds LPL in the subendothelial space and transports it to the luminal surface of endothelial cells, where it anchors the LPL protein by means of the GPI domain. Genetic studies have shown that mutations in GPIHBP1, and mutations in LPL that make it unable to bind GPIHBP1, are both associated with increased plasma triglyceride concentrations.

Angiopoietin-like protein 3

Angiopoietin-like protein 3 (ANGPTL3) acts in the liver to promote the production and secretion of apo B containing lipoproteins. In the circulation, it inhibits the activities of LPL. It may also affect the activity of EL and thereby modulate HDL cholesterol concentrations. Loss of function mutations in the *ANGPTL3* gene have been associated with a combined hypolipidaemia phenotype, with low plasma concentrations of cholesterol, HDL-cholesterol and triglycerides.

CLASSIFICATION OF LIPOPROTEIN DISORDERS

Fredrickson originally suggested dividing the hyperlipidaemias into five types (I–V) based on which of the apo B-containing lipoprotein classes were shown to be increased on paper electrophoresis of a fasting sample.

TABLE 37.4 World Health Organization classification of hyperlipidaemia

Type	Major lipid abnormality	Minor lipid abnormality	Electrophoretic mobility	Lipoprotein abnormality
I	↑ Triglycerides	↑ Cholesterol	Staining at origin	Chylomicrons present
IIa	↑ Cholesterol		↑ β band	↑ LDL
IIb	↑ Cholesterol ↑ Triglycerides		↑ β and pre-β band	↑ LDL and ↑ VLDL
III	↑ Cholesterol ↑ Triglycerides		Broad β band	IDL present in detectable amounts
IV	↑ Triglycerides	↑ Cholesterol	↑ Pre-β band	↑ VLDL
V	↑ Triglycerides	↑ Cholesterol	↑ Pre-β band and staining at origin	↑ VLDL and chylomicrons present

This was subsequently modified by the World Health Organization, which introduced a subdivision of type II into two (types IIa and IIb) (Table 37.4).

The Fredrickson classification is limited in being only a description of the lipoprotein phenotype manifested in an individual, and only covers abnormal elevations of the apo B containing lipoproteins. It does not differentiate primary from secondary causes of any particular phenotype, and does not include deficiencies of the apo B containing lipoproteins or disorders of HDL metabolism. It soon became clear that the type I phenotype can be caused by any one of three inherited abnormalities, all of which result in a functional deficiency of lipoprotein lipase. Type III is caused by an inherited defect in apo E causing reduced binding to the LRP (apo E) receptor in the liver. The other types (IIa, IIb, IV and V) can each be a consequence of

one or more inherited defects or may reflect secondary causes. Thus classifying an individual into one of these Fredrickson classes gives no indication as to the genetic basis of the manifested lipoprotein disorder, its mechanism, the need for family screening or the management required. It is now recognized that even the type I and type III phenotypes can have multiple other genetic causes due either to single gene disorders or the cumulative effect of multiple small genetic variants together with secondary environmental influences. For all these reasons, this classification has largely been superseded by a clinical and genetic classification, which relates to the molecular mechanism underlying the clinical phenotype. The genetic basis of the primary dyslipidaemias and the relationship, where known, to the corresponding clinical and Fredrickson phenotypes is given in Table 37.5.

TABLE 37.5 Clinical/genetic classification of dyslipidaemias

Clinical disorder	Lipid changes			WHO phenotype	Lipoprotein	Protein	Gene
	LDL	TG	HDL				
Dysbetalipoproteinaemias							
Decreased beta lipoproteins							
Abetalipoproteinaemia	↓↓	↓↓	↓	N/A	Absent Apo B	MTP	<i>MTP</i>
Chylomicron retention disease	↓	→	↓	N/A	Absent B-48, decreased B-100	Srb1	<i>SARA2</i>
Familial hypobetalipoproteinaemia	↓	↓	↓	N/A	Decreased Apo B	Apo B	<i>APOB</i>
Increased beta lipoproteins							
Familial combined hyperlipidaemia	↑	↑	→/↓	IIb/IV	VLDL		<i>USF1?</i>
Familial hypertriglyceridaemia							
LPL deficiency	↑	↑↑↑	↓	I	Chylo	LPL	<i>LPL</i>
Apo C-II deficiency	↑	↑↑↑	↓	I	Chylo	Apo CII	<i>APOC2</i>
LPL inhibitor	↑	↑↑↑	↓	I	Chylo	LPL inhibitor	
Combined lipase deficiency	↑/→	↑↑	→/↓	I	Chylo	EL/HL/LPL	<i>LMF1</i>
Not defined	↑/→	↑↑	→/↓	I/III/IV/V	Chylo, VLDL, IDL		Loss of function mutations in single genes (e.g. <i>GPIHBP1, APOA5</i>) or cumulative small effect variants in multiple genes (e.g. <i>APOA5; LPL; GCKR; APOB</i>)
Type III (remnant) hyperlipidaemia	↑↑	↑↑	→/↓	III	IDL	Apo E-2	<i>APOE</i>

TABLE 37.5 Clinical/genetic classification of dyslipidaemias (continued)

Clinical disorder	Lipid changes			WHO phenotype	Lipoprotein	Protein	Gene
	LDL	TG	HDL				
Familial hypercholesterolaemia							
Autosomal dominant							
Classic FH	↑↑	→	→	IIa	LDL	LDLR	<i>LDLR</i>
Familial defective apo B	↑	→	→	IIa	LDL	ApoB	<i>APOB</i>
Gain of function PCSK9	↑↑↑	→	→	IIa	LDL	Increased PCSK9	<i>PCSK9^a</i>
Autosomal recessive							
Autosomal recessive hypercholesterolaemia	↑↑↑	→	→	IIa	LDL	LDLRAP1	<i>LDLRAP1</i>
Dysalphalipoproteinaemia							
Decreased alpha lipoproteins							
Tangier disease	↓	→/↑	↓↓	N/A		ABCA1	<i>ABCA1</i>
Apo A-I deficiency	→	→/↓	↓↓↓	N/A		Apo A-I	<i>APOA1</i>
Apo A-I variants	→	→	↓	N/A		e.g. Apo A-I Milano	<i>APOA1</i>
LCAT deficiency	↓	↑	↓	N/A		LCAT	<i>LCAT</i>
Fish eye disease (partial LCAT deficiency)	↓	↑	↓	N/A		LCAT	<i>LCAT</i>
Not defined			↓	N/A		EL	<i>LIPG^a</i>
Increased alpha lipoproteins							
Hepatic lipase deficiency	↑	↑↑	↑	N/A	HDL	HL	<i>LIPC</i>
CETP deficiency	↓	→	↑↑	N/A	HDL	CETP	<i>CETP</i>
Not defined			↑	N/A	HDL	SRB1	<i>SCARB1</i>
Not defined			↑	N/A	HDL	EL	<i>LIPG</i>

^aDenotes gain of function mutation.

Lipids may be deposited in certain extravascular sites in the presence of significant hyperlipidaemia. In certain instances, the site and form of this lipid deposition may be characteristic of the underlying lipid abnormality (Table 37.6).

THE PRIMARY DYSLIPOPROTEINAEMIAS

A primary dyslipoproteinaemia is an inherited disorder of lipoprotein metabolism that may manifest as hyperlipidaemia, hypolipidaemia or normolipidaemia associated with lipoproteins of abnormal composition

or an abnormal distribution of the normal lipoprotein classes.

Hypobetalipoproteinaemia

The term 'hypobetalipoproteinaemia' describes the situation when the total plasma cholesterol, LDL-cholesterol or apo B concentrations are less than the 5th centile. This may be secondary to an underlying disorder, for example fat malabsorption, or a consequence of an inherited defect in lipoprotein metabolism (primary hypobetalipoproteinaemia). Primary hypobetalipoproteinaemia encompasses three disorders: abetalipoproteinaemia, chylomicron retention disease and familial hypobetalipoproteinaemia.

TABLE 37.6 Extravascular manifestations of hyperlipidaemia

Extravascular site of lipid deposition	Type of hyperlipidaemia
Xanthelasma (periorbital)	Any form of hypercholesterolaemia; may occur with normal lipid concentrations; common in hyperlipidaemias due to cholestasis
Corneal arcus (around the iris of the eye)	Arcus senilis can occur with increasing age in subjects with normal lipid concentrations; if present in a subject <45 years, probably denotes presence of significant underlying hyperlipidaemia (usually ↑ LDL)
Tendon xanthomas (occur in extensor tendons of hands, feet, Achilles tendon and patellar tendon)	Virtually pathognomonic of familial hypercholesterolaemia; also found in the very rare conditions, β-sitosterolaemia and cerebrotendinous xanthomatosis
Tuberous and tuberoeruptive xanthomas (occur over extensor surfaces of limbs, i.e. elbows/knees)	Remnant hyperlipoproteinaemia and familial hypercholesterolaemia
Palmar crease xanthomas	Remnant hyperlipidaemia
Eruptive xanthomas (occur in crops, particularly over the extensor surface of the elbows and over the buttocks)	Chylomicronaemia

Abetalipoproteinaemia

This is a very rare autosomal recessive condition, first described in 1950 by Bassen and Kornzweig. It typically presents in infancy with failure to thrive and chronic diarrhoea owing to fat malabsorption. Other features include acanthocytosis (caused by abnormal erythrocyte membrane lipid composition) and an atypical retinitis pigmentosa (as a result of vitamin A deficiency). Later manifestations are night blindness and neurological disability, particularly ataxia, caused by vitamin A and vitamin E deficiencies, respectively. The primary aim of treatment in these subjects is to ensure that fat-soluble vitamin intake is sufficient, which may require their parenteral administration.

There is a complete absence of all apo B-containing lipoproteins in the plasma, and no circulating apo B, whereas HDL and apo A-I concentrations are approximately 50% of normal. Plasma cholesterol concentration is low (typically 0.5–1.5 mmol/L) and plasma triglyceride concentration is very low (<0.2 mmol/L). On microscopy, lipid droplets can be seen to accumulate intracellularly in hepatocytes and enterocytes, suggesting a defect in the assembly of the apo B containing lipoproteins. The cause is now known to be loss of function mutations in the *MTP* gene (see p. 720). In the presence of defective MTP, no apo B-containing lipoproteins are formed. The malabsorption of fat and the fat-soluble vitamins is a result of the failure of chylomicron formation.

Chylomicron retention disease

The inheritance of chylomicron retention disease is probably autosomal recessive. It is characterized by an absence of apo B-48 in plasma and lack of a postprandial lipaemic response. Affected individuals also have low LDL, HDL and fat-soluble vitamin concentrations and their LDLs are enriched in triglycerides. They have fat malabsorption and steatorrhoea and, without supplementation of fat-soluble vitamins, develop neurological dysfunction. Enterocytes of affected individuals contain fat droplets, and apo B-48 is demonstrable immunochemically within the enterocytes. Chylomicron retention disease is caused by mutations in the *SARA2* gene. The protein product of the *SARA2* gene is sar1b, which is involved in the transport of chylomicrons through the enterocyte secretory pathway (see p. 712).

Familial hypobetalipoproteinaemia

Familial hypobetalipoproteinaemia (FHBL) is an autosomal co-dominant disorder resulting from mutations in *APOB* generating premature stop codons and therefore truncation of apo B. These truncated forms have a reduced capacity for lipid binding and thus result in the secretion of smaller, denser, relatively lipid-poor lipoprotein particles. Lipoprotein particles containing the larger truncated forms (e.g. apo B-89 and apo B-75) fall into the VLDL fraction when secreted, whereas those containing the smaller forms (e.g. apo B-29) fall into the HDL fraction. There is a threshold in apo B size under which the apo B cannot form a lipoprotein; this threshold is between apo B-28 and apo B-29. Mutations producing truncated forms of apo B greater than apo B-48 are associated with

the production of normal chylomicrons. Apolipoprotein B forms smaller than B-67 are incorporated into VLDL in the normal way but are not able to interact with LDL receptors.

Individuals heterozygous for FHBL associated with truncated apo Bs are often asymptomatic, but a small proportion have loose stools. The plasma LDL-cholesterol concentration in heterozygotes would be expected to be about 50% of that in unaffected family members, but concentrations of about a third of normal are actually observed, possibly because of reduced hepatic secretion or upregulation of LDL receptors that results in increased clearance. A high incidence of non-alcoholic fatty liver disease is reported in FHBL heterozygotes. Homozygous FHBL clinically resembles abetalipoproteinaemia.

Familial combined hyperlipidaemia

This condition is the most common lipid disorder in subjects presenting with ischaemic heart disease. It is inherited as an autosomal dominant trait, but usually does not become manifest until adulthood. The typical findings are a raised plasma apo B-100 concentration and an increase in plasma concentrations of LDL (high total cholesterol), VLDL (high triglycerides) or both. Different individuals within a kindred may have any of these phenotypes, and in a single individual, the phenotypic pattern may change over time. Plasma HDL-cholesterol concentration is usually low; this probably results from increased transfer of cholesteryl esters from HDL to the triglyceride-rich lipoproteins. The LDL particles in familial combined hyperlipidaemia (FCH) tend to be smaller and denser than usual; thus individuals with FCH display the 'atherogenic lipid profile' comprising high plasma triglyceride and low HDL-cholesterol concentrations and small, dense LDL particles.

Overproduction of apo B-100 occurs in FCH, so that affected individuals have raised plasma apo B concentrations even though their lipid concentrations may be normal. The development of hyperlipidaemia is dependent on an increase in the availability of hepatic triglycerides; this explains the observation that affected subjects are frequently obese, and may have other features of the metabolic syndrome. The phenotype expressed is thought to reflect the efficiency with which the VLDL is processed in affected individuals. Thus individuals who have one abnormal LPL gene (parents of children with LPL deficiency) may appear to have FCH because of the resultant slowed clearance of VLDL.

Genetic studies have suggested a link between FCH and the upstream transcription factor 1 (USF1) gene. This is a transcription factor responsible for upregulating the transcription of a number of genes involved in glucose and lipid metabolism, including those coding for apo A-V.

Familial hypertriglyceridaemia

Some families show Mendelian inheritance of hypertriglyceridaemia. In most cases this reflects an increase in VLDL, with moderately elevated triglyceride concentrations (4–10 mmol/L). In some cases, there is a more severe hyperlipidaemia (>10 mmol/L) and fasting

chylomicronaemia. The latter causes a Fredrickson type I phenotype, while less severe hypertriglyceridaemia generally causes a type IIb, IV or V pattern. Severe hypertriglyceridaemia is typically autosomal recessive and in addition to defects in LPL and apo C-II, which have long been recognized as causes of type I hyperlipidaemia (chylomicronaemia syndrome), it is now recognized that loss of function mutations in genes including *APOA5*, *LMF1* and *GPIHBP1* may produce a similar phenotype, often inherited in an autosomal recessive manner. By contrast, the cumulative effect of multiple small effect polymorphisms in genes including *LPL*, *APOA5*, *GCKR* and *APOB*, interact with environmental factors to produce more moderate hypertriglyceridaemia, and a variety of lipoprotein phenotypes. These include IIb, IV, V and, sometimes, type III. The environmental factors interacting with these genetic variants to produce moderate or severe fasting hypertriglyceridaemia are the same as those causing secondary hyperlipidaemia, e.g. diabetes, insulin resistance or excessive alcohol consumption. VLDL particles are larger than usual, and relatively deficient in apo B in familial hypertriglyceridaemia, which suggests that an important factor is overproduction of triglycerides by the liver. Chylomicronaemia syndrome, caused by autosomal recessive major loss of function mutations in single genes, often manifests in childhood, whilst other causes of familial hypertriglyceridaemia rarely manifest before adulthood.

In patients with a fasting plasma triglyceride concentration >10 mmol/L, there is a risk of acute pancreatitis, and reducing this risk is usually the first priority of treatment, before addressing residual cardiovascular risk. Fibrates or ω 3-fatty acids, in combination with a low fat diet and avoidance of excessive alcohol consumption, are usually effective. Plasma exchange, and intravenous insulin infusion (to increase lipoprotein lipase activity), have both been used in severe cases although the evidence for their long-term effectiveness is limited.

Chylomicronaemia syndrome

Chylomicronaemia syndrome manifests as eruptive xanthomata (see Fig. 37.11), lipaemia retinalis, hepatosplenomegaly and recurrent bouts of abdominal pain that may be symptoms of acute pancreatitis. Plasma triglyceride concentrations are markedly increased (>10 mmol/L), HDL-cholesterol concentrations are very low and chylomicrons are present in the fasting state. Three inherited causes of this condition have been described: lipoprotein lipase deficiency, apo C-II deficiency and familial lipoprotein lipase inhibitor. Secondary causes, including excessive alcohol consumption and newly diagnosed type 1 diabetes, may produce a very similar phenotype in genetically predisposed individuals.

Lipoprotein lipase deficiency. This is a rare autosomal recessive condition with an incidence of approximately 1 in 1 000 000. Presentation is usually in childhood with recurrent abdominal pain. Affected individuals may have low, normal or increased concentrations of immunoreactive LPL in post-heparin plasma but catalytic activity is undetectable. Gene replacement therapies for

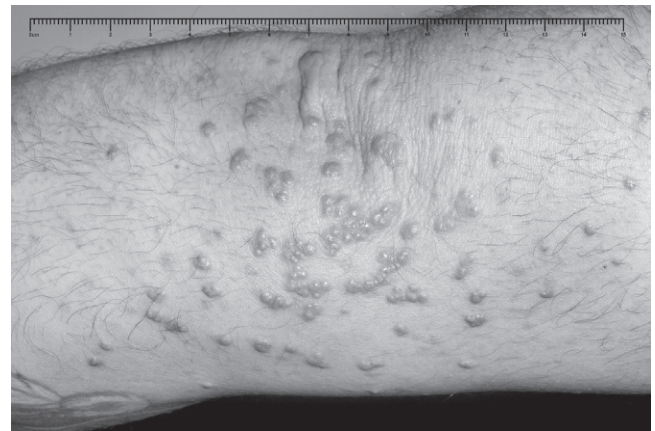


FIGURE 37.11 ■ Eruptive xanthomata over the elbow.

the most severe forms of this condition are undergoing clinical trials.

Obligate heterozygotes for LPL deficiency have been shown to have low plasma LPL activity: some may develop severe hypertriglyceridaemia, and a small number, pancreatitis.

Apo C-II deficiency. Homozygous apo C-II deficiency tends to present later and be somewhat milder than LPL deficiency. Apo C-II is detectable in about 50% of individuals with apo C-II deficiency, but the protein produced is unable to activate LPL. It is possible that the concentrations of the triglyceride-rich lipoproteins may reach a level at which LPL can hydrolyse them even in the absence of apo C-II: this could account for the milder form of the disease as compared with that seen with primary deficiency of the enzyme.

Relatives who are obligate heterozygotes for the abnormal gene usually have normal lipid and lipoprotein concentrations. This implies that up to a 50% reduction in the apo C-II concentration does not compromise the rate of chylomicron and VLDL clearance.

Familial lipoprotein lipase inhibitor. Families have been described whose affected members have chylomicronaemia apparently caused by an inhibitor of LPL. The defect appears to be inherited in an autosomal dominant fashion. The inhibitor has not been identified.

Remnant hyperlipoproteinaemia

The characteristic feature of remnant hyperlipoproteinaemia (RH) is cholesterol enrichment of the VLDL fraction, reflecting an accumulation of remnant particles, both chylomicron remnants and intermediate density lipoprotein. These remnants are responsible for the abnormal electrophoretic pattern, which gave the condition its former name of broad β disease (type III in the Fredrickson classification), since a broad β band is seen on electrophoresis instead of the usual two distinct pre- β (VLDL) and β (LDL) bands. Remnant hyperlipidaemia is also sometimes referred to as dysbetalipoproteinaemia. Subjects with this disorder show a mixed hyperlipidaemia, often with approximately equal plasma concentrations of cholesterol and triglycerides.

The predominant defect found in most subjects with the condition is the presence of apo E2, which displays a recessive mode of inheritance. While apo E2 homozygosity is essential for the accumulation of remnant particles, it is not sufficient for the manifestation of the condition. A super-imposed genetic or environmental factor is necessary before remnant hyperlipidaemia becomes manifest, such as the concomitant inheritance of another primary hyperlipidaemia or, more frequently, a secondary cause such as obesity, excessive alcohol consumption, diabetes or hypothyroidism. The 10% of individuals with RH who do not show homozygosity for apo E2/2 have a variety of molecular defects, including point mutations resulting in amino acid substitutions at residues 142, 145 or 146, or the insertional variant, apo E-Leiden; a few lack apo E entirely.

Remnant hyperlipoproteinaemia rarely presents before adulthood. It is associated with characteristic xanthomata occurring in the palmar creases (Fig. 37.12) which are virtually pathognomonic, and tuberous or tuberoeruptive xanthomata over the extensor surfaces of the elbows and knees (Fig. 37.13). There is an



FIGURE 37.12 ■ Xanthomata in the palmar creases.



FIGURE 37.13 ■ Tuberous xanthomata.

increased risk of premature atherosclerosis involving the peripheral vascular system as well as the coronary arteries.

Familial hypercholesterolaemia

This term encompasses a group of disorders due to mutations causing reduced clearance of LDL by the LDL receptor resulting in marked hypercholesterolaemia and premature atherosclerosis. 'Classic' familial hypercholesterolaemia (FH) is due to mutations in the *LDLR* gene. Autosomal dominant mutations are now recognized in three genes causing conditions collectively referred to as autosomal dominant hypercholesterolaemia (ADH), which includes classic FH. A further gene locus has recently been associated with this clinical phenotype, but the causative gene has not yet been identified. A rare autosomal recessive condition has also been described with a similar clinical and biochemical phenotype. The four clearly defined disorders that result in FH are as follows:

- classic FH – ADH 1
- familial defective apolipoprotein B-100 – ADH 2
- gain of function mutation in PCSK9 – ADH 3
- loss of function mutation in LDLRAP1 – autosomal recessive hypercholesterolaemia.

Classic familial hypercholesterolaemia (FH). This is the most severe of the hyperlipidaemias with respect to the propensity of affected, untreated individuals to develop atherosclerosis. It manifests in the heterozygous state as marked hypercholesterolaemia, owing to high LDL-cholesterol concentration, and premature cardiovascular disease. Heterozygous FH is one of the commonest inherited metabolic diseases, with a frequency of about 1 in 500 in most populations. Tendon xanthomata in the extensor tendons of the digits (Fig. 37.14) and in the Achilles tendons (Fig. 37.15) occur in approximately 70% of untreated heterozygotes and their presence in affected individuals or first-degree relatives aids the clinical diagnosis, which is based on criteria defined by the Simon Broome Trust (Box 37.1).

Based on the frequency of the heterozygous state, homozygous FH would be expected to have an incidence of one in 1 000 000. Most apparently homozygous FH



FIGURE 37.14 ■ Xanthomata in the extensor tendons of the fingers.



FIGURE 37.15 ■ Xanthomata in the Achilles tendon.

BOX 37.1 Diagnostic criteria for familial hypercholesterolaemia (FH)

A diagnosis of definite FH using the Simon Broome^a criteria requires:

- serum cholesterol >7.5 mmol/L (>6.7 mmol/L in children <16 years) *or* LDL cholesterol >4.9 mmol/L (>4.0 mmol/L in children <16 years) *plus*
 - tendon xanthomas in the patient or first- or second-degree relative
- or*
- DNA-based evidence of a pathogenic mutation in *LDLR*, *APOB* or *PCSK9*

A diagnosis of possible FH requires:

- serum cholesterol concentration >7.5 mmol/L (>6.7 mmol/L in children <16 years) *or* LDL cholesterol >4.9 mmol/L (>4.0 mmol/L in children <16 years) *plus*
 - a family history of myocardial infarction before age 60 years in a first-degree relative or before age 50 years in a second-degree relative
- or*
- serum cholesterol concentration >7.5 mmol/L (6.7 mmol/L in children <16 years) in a first- or second-degree relative.

^aThe Simon Broome Trust maintains a register of FH patients in the UK. It is named after a patient with the condition.

subjects are actually compound heterozygotes. The hypercholesterolaemia is far more severe in subjects with two mutant genes; cutaneous xanthomata often occur in childhood and coronary artery disease may present during the first decade of life. LDL apheresis has now become standard treatment in these patients in addition to lipid lowering drugs and diet. This can significantly increase life expectancy, particularly if the condition is diagnosed at a young age.

Classic FH, which accounts for the defect in 80–90% of patients with familial hypercholesterolaemia, is due to mutations of the LDL receptor gene. Around 20 common

mutations account for 50% of cases, but over 1200 mutations have now been described, including a number of pathogenic intron mutations.

In vitro, it has been demonstrated that cells from FH heterozygotes have about half the number of functional LDL receptors than do cells from normal individuals. The effect of this deficiency of functional LDL receptors in vivo is that plasma LDL-cholesterol concentrations are increased to approximately twice the usual level, owing to a combination of reduced LDL uptake and LDL overproduction. The overproduction of LDL results from defective hepatic IDL uptake by the LDL receptor so that extracellular IDL to LDL conversion is increased. Hepatic cholesterol synthesis is also upregulated because of reduced LDLR-mediated cholesterol uptake.

Familial defective apolipoprotein B-100 (FDB). This is an autosomal co-dominant disorder that results from an abnormality in the LDLR-binding domain of apo B-100. Most cases are due to mutations in the *APOB* gene at the codon for amino acid 3500 where Arg to Gln and Arg to Trp substitutions have been described. The arginine to glutamine substitution has been shown to change the conformation of the C-terminal tail, which results in reduced binding to the LDLR. Other mutations have been described that have a less severe effect on the ability of apo B to bind to LDL receptors. FDB is phenotypically very similar to classic FH; affected subjects have elevated plasma total and LDL cholesterol concentrations but the hypercholesterolaemia is usually less severe and tendon xanthomas are less common.

Gain of function mutation in PCSK9. This variant of FH was characterized by studying families that phenotypically were similar to classic FH and FDB. It was shown to be caused by gain of function mutations in the gene coding for proprotein convertase subtilisin kexin 9 (PCSK9). These mutations increase the rate at which PCSK9 degrades LDL receptors, resulting in reduced receptor expression on the cell surface. These mutations tend to be associated with a more severe clinical phenotype, with affected individuals showing higher plasma LDL-cholesterol concentrations and more aggressive vascular disease than seen in classic FH.

Autosomal recessive hypercholesterolaemia. Patients with autosomal recessive hypercholesterolaemia (ARH) phenotypically resemble classic FH homozygotes. They have large xanthomas (tendon, tuberous, planar) that present in childhood, and develop premature atherosclerosis, particularly of the coronary and carotid arteries. They also develop aortic stenosis. Plasma LDL-cholesterol concentrations in patients with ARH are generally higher than those found in classic FH but not as high as those seen in homozygous FH.

Autosomal recessive hypercholesterolaemia results from a mutation in the *LDLRAP1* gene. LDLRAP1 mediates normal LDL receptor-mediated endocytosis after LDL is bound to the receptor, and this process is defective in ARH. Obligate heterozygotes for ARH

who carry a single pathogenic *LDLRAP1* mutation have normal plasma LDL-cholesterol concentrations.

Polygenic hypercholesterolaemia

Hypercholesterolaemia is more common in Western societies than would be expected, given the frequency of the individual monogenic disorders discussed above. The remainder of individuals with primary hypercholesterolaemia are therefore referred to as having polygenic hypercholesterolaemia. The term 'polygenic' is used because the individuals' genetic background affects the extent to which lipoprotein metabolism will be influenced by environmental factors.

The monogenic disorders are associated with the same increased risk for atherosclerosis in different populations. However, the frequency with which polygenic hypercholesterolaemia occurs in different populations varies widely, as does the incidence of coronary artery disease. Populations with a high mean plasma cholesterol concentration have a high rate of coronary artery disease. It is believed that the difference between populations in the frequency of hypercholesterolaemia reflects one or more environmental factors, the most obvious being diet and, in particular, the saturated fat content of the diet.

The frequency with which polygenic hypercholesterolaemia is diagnosed in a population depends on the value taken as the upper cut-off for a 'normal' cholesterol. If 5 mmol/L is used, about 50% of most adult populations will be classified as hypercholesterolaemic, as the population mean cholesterol concentration is typically 5.5–6.0 mmol/L in many Western countries. At this level, CVD risk is already twice that of a population with a mean cholesterol <4.0 mmol/L, so a statistical definition of 'normal' based on population values is unhelpful. Laboratories will generally refer to 'ideal' cholesterol targets. In the UK, these are currently total cholesterol <4.0 mmol/L and LDL-cholesterol <2.0 mmol/L.

Dysalphalipoproteinaemias

The dysalphalipoproteinaemias are disorders of the synthesis and secretion of apo A-containing lipoproteins. With the exception of apo A-I Milano (see below), all other genetic causes of familial hypoalphalipoproteinaemia are associated with some degree of increased cardiovascular risk.

Abnormal apolipoprotein A structure

Population screening for apo A-I structural variation has revealed at least 11 variants. The best described is apo A-I Milano, which is associated with a low HDL (HDL2) concentration but no increased incidence of atherosclerosis. Apolipoprotein A-I Milano results from substitution of cysteine for arginine at position 173, which has the effect of changing the physical property of one of the amphipathic helical regions involved in lipid binding and also allows disulphide bonding to other proteins. Dimers with apo A-II and apo E have been observed. In addition, LCAT activity is reduced.

Apo A-I deficiency

At least three types of apo A-I deficiency have been described, all associated with corneal clouding and premature coronary heart disease. Heterozygotes have 50% of normal HDL concentrations. HDL is virtually undetectable in homozygotes. In one type there is deficiency only of apo A-I, another also has deficiency of apo C-III and has low plasma triglyceride concentrations. The *APOC3* gene is 3' to the *APOA1* gene but in the opposite orientation. These two genes are normally convergently transcribed from opposite DNA strands. In patients with combined deficiency of apo A-I and apo C-III, an inversion of approximately 5.5 kb containing portions of the *APOA1* and *APOC3* genes has been identified. In a third form, there is deficiency of apo A-I, apo C-II and also apo A-IV. These may show fat malabsorption due to the deficiency of apo A-IV.

Disorders of HDL metabolism (Box 37.2)

Tangier disease

Tangier disease is a rare autosomal recessive condition characterized by storage of cholesteryl esters in macrophages. This is responsible for the characteristic orange appearance of the tonsils. Splenomegaly also occurs, and is often accompanied by mild thrombocytopenia and reticulocytosis, but hepatomegaly and lymphadenopathy are less constant features. Corneal clouding may occur and most affected individuals have some neurological dysfunction.

Total plasma cholesterol concentrations are low (typically <3 mmol/L) but, unlike the apo B deficiency states, plasma triglyceride concentrations are normal or increased. Virtually no HDL-cholesterol is present (<0.1 mmol/L) and only pre- β_1 HDL particles are detectable in plasma. The concentrations of apo A-I and apo A-II are approximately 1–3% and 5%, respectively, of values found in normal individuals. An increase in pro-apo A-I can be demonstrated by isoelectric focusing but rapid catabolism of pro-apo A-I results in reduced conversion of pro-apo A-I to the mature form and consequently low plasma concentrations of apo A-I. Although it appears that Tangier patients are at increased cardiac risk, the level of risk varies between affected families and

BOX 37.2 Inherited disorders affecting HDL cholesterol

Increased HDL cholesterol

- Hepatic lipase deficiency^a
- CETP deficiency

Reduced HDL cholesterol

- Apo A-I deficiency^a
- Apo A-I variants, e.g. apo A-I Milano
- LPL deficiency
- Apo C-II deficiency^a
- Tangier disease
- Familial LCAT deficiency^a
- Fish eye disease^a

^aAssociated with premature atherosclerosis.

is not as high as predicted from epidemiological studies based on their plasma HDL-cholesterol concentrations. This may be related to their low LDL-cholesterol concentrations, but thrombocytopenia and relative platelet hyporeactivity may also be important factors.

Tangier disease is caused by the loss of function of the cholesterol-efflux regulatory protein, ABCA1, resulting from mutations in the *ABCA1* gene. Individuals homozygous for *ABCA1* mutations are at increased risk of coronary artery disease compared with unaffected individuals and heterozygous individuals are at intermediate risk.

Familial lecithin-cholesterol acyltransferase deficiency

Familial LCAT deficiency is an autosomal recessive condition. The gene for LCAT is located on chromosome 16; mutations in this gene that are associated with no LCAT activity result in familial LCAT deficiency. The clinical features of this condition include corneal opacities, haemolytic anaemia, proteinuria, high plasma triglyceride and low HDL-cholesterol concentrations. Despite the low HDL-cholesterol concentrations, premature atherosclerosis is rarely seen. Heterozygous carriers lack clinical features but have HDL-cholesterol concentrations about 50% of normal.

Lecithin-cholesterol acyltransferase deficiency results in an inability to esterify free cholesterol in plasma and hence the accumulation of free cholesterol in all lipoprotein fractions. Lipoprotein particles may also be of abnormal size or shape as well as composition; thus the HDL particles are either disc shaped or unusually small spherical particles similar to newly secreted HDL. Lipoprotein X is detectable on electrophoresis.

Fish eye disease

Fish eye disease is an autosomal recessive disorder resulting from mutations in the *LCAT* gene. In contrast to familial LCAT deficiency, there is detectable LCAT activity. Corneal opacities that give the eyes the appearance of those of dead fish have been described in association with low HDL-cholesterol concentrations (approximately 10% of normal). LDL is enriched in triglyceride; the concentration of VLDL is increased but the composition is normal.

Hepatic triglyceride lipase deficiency

Several familial cases of HL deficiency have been reported and inheritance appears to be autosomal recessive, with heterozygotes showing an intermediate phenotype. Homozygotes have severe hypertriglyceridaemia, hypercholesterolaemia, and increased IDL concentrations, producing a phenotype similar to type III or remnant hyperlipidaemia. HL deficiency is distinguished by an absence of post heparin HL activity. Apolipoprotein A-I and HDL-cholesterol concentrations are slightly raised, with HDL particles being abnormally triglyceride-rich HDL2. Premature vascular disease has been reported in HL deficiency, although the exact role of HL in the development of atherosclerosis remains uncertain, as increased HL activity (e.g. secondary to anabolic steroid use) may also result in increased vascular disease.

Cholesterol ester transfer protein deficiency

Several mutations of the *CETP* gene that cause CETP deficiency have been described. Homozygous CETP-deficient subjects have high plasma concentrations of HDL-cholesterol, apo A-I, apo A-II and apo E. Plasma HDL-cholesterol concentrations are usually >3 mmol/L in homozygous subjects who have a complete absence of CETP activity, values of 2–2.5 mmol/L are typically seen in heterozygotes. These subjects also have an approximately 40% reduction in LDL-cholesterol and apo B. Cholesterol ester transfer protein deficiency is relatively common in Japan, where it accounts for about half of all cases of hyperalphalipoproteinaemia, but it is rare in Caucasians. There is evidence that, when associated with a high plasma HDL concentration, CETP deficiency protects against the development of atheroma. On this basis, drugs that inhibit CETP have been developed: these produce an increase in plasma HDL-cholesterol concentration of over 100% and fall in LDL-cholesterol concentration of around 40%. Because of the complex relationship between CETP activity, HDL concentrations and atherosclerosis, the effect of these agents on cardiovascular outcomes will need to be tested in clinical trials before it is clear whether or not CETP inhibition has therapeutic value.

ACQUIRED HYPERLIPIDAEMIAS

Acquired, or secondary, hyperlipidaemia is caused by alterations in lipoprotein metabolism resulting from another disease state or from drug therapy. Treatment of the underlying condition or discontinuation of the precipitating drug may correct the hyperlipidaemia. However, in some cases, such as chronic kidney disease or diabetes, where metabolism remains disturbed despite treatment, hyperlipidaemia may persist. In these conditions the increased vascular risk is a function of qualitative as well as quantitative changes in plasma lipoproteins. In particular, LDL exists in a small, dense form which is more readily oxidized and consequently more atherogenic. Cardiovascular risk therefore remains higher than is predicted from cholesterol and triglyceride measurements alone. The secondary hyperlipidaemias are summarized in [Box 37.3](#).

Xanthomata may occur in acquired hyperlipidaemia just as in the primary lipid disorders with the same pattern of lipid abnormalities: thus eruptive xanthomata occur in the chylomicronaemia syndrome, whatever its aetiology. Florid cutaneous lipid deposition may occur in the presence of abnormal lipoproteins such as lipoprotein X in cholestasis (see below) or where monoclonal immunoglobulins bind lipoproteins or their receptors and interfere with lipid metabolism, as may occur in monoclonal gammopathy of uncertain significance (MGUS) and multiple myeloma.

Diabetes mellitus

Poorly controlled diabetes gives rise to hypertriglyceridaemia. In both types of diabetes, there is insulin deficiency,

BOX 37.3 The more common causes of secondary hyperlipidaemia**Hypercholesterolaemia (↑ LDL^a)**

- Hypothyroidism
- Nephrotic syndrome
- Immunoglobulins^b
- Cholestasis
- Anorexia nervosa
- Drugs
 - Ciclosporin
 - Sirolimus
 - Antiepileptics (↑ HDL also)

Hypertriglyceridaemia^c (↑ VLDL)

- Diabetes mellitus
- Chronic kidney disease
- Obesity
- Alcohol
- Hypothyroidism
- Immunoglobulins^b
- Drugs
 - Oestrogens
 - Corticosteroids
 - Progestogens
 - Protease inhibitors
 - Retinoids
 - Second generation antipsychotics

Mixed hyperlipidaemia (↑ VLDL, ↑ LDL)

- Hypothyroidism
- Nephrotic syndrome
- Immunoglobulins^b

^aExcept cholestasis in which there is an increase in the abnormal lipoprotein, lipoprotein X.

^bThe presence of abnormal immunoglobulins, as occurs in multiple myeloma, benign monoclonal gammopathies and systemic lupus erythematosus can give rise to any abnormality of the lipoprotein pattern due to binding to lipoproteins, enzymes or receptors and thereby preventing normal lipoprotein metabolism.

^cOften accompanied by a low HDL.

either absolute in type 1 diabetes or relative in type 2 diabetes. Insulin activates lipoprotein lipase and thereby enhances clearance of triglyceride-rich lipoproteins, but has the opposite effect on the HSL of adipose tissue. Thus, in insulin deficiency there is, in addition to reduced clearance, an increased influx of free fatty acids to the liver, leading to increased hepatic triglyceride synthesis.

In well-controlled diabetes, although plasma total cholesterol and triglyceride concentrations may be normal, there often remains a significant dyslipidaemia, sometimes referred to as an 'atherogenic lipoprotein phenotype'. Plasma HDL-cholesterol concentration is usually low (<1.1 mmol/L) and triglyceride concentration elevated (>1.7 mmol/L). This pattern is indicative of the presence of small, dense LDL particles, which are more susceptible to oxidation and more atherogenic. It has been demonstrated that apolipoproteins A-I, A-II, B, C-I and E become glycosylated in patients with diabetes: it is possible that this glycation affects the normal uptake of remnant particles, resulting in their persistence in the circulation with atherogenic consequences.

The genetic background on which a secondary hyperlipidaemia is superimposed will affect its severity; for example, diabetic patients who manifest fasting chylomicronaemia are likely to have an underlying primary hypertriglyceridaemia.

Hypothyroidism

Various lipid abnormalities may occur in patients with untreated hypothyroidism, but the commonest is an increase in plasma LDL-cholesterol concentration. This is a consequence of a reduction in LDL receptors, resulting in reduced clearance of LDL. Lipoprotein lipase activity may also be impaired in hypothyroidism, which explains the hypertriglyceridaemia that sometimes occurs.

Since both overt and subclinical hypothyroidism are relatively common, it is imperative that all subjects found to have hyperlipidaemia are screened for hypothyroidism. If they are found to be hypothyroid, treatment with thyroxine should be instituted. If the subject does not have cardiovascular disease, the lipid profile should be rechecked once the euthyroid state is regained before a decision on lipid-lowering treatment is made. If the subject already has cardiovascular disease when hypothyroidism is diagnosed, it may be unwise to delay lipid lowering therapy, but it should be instituted with care since hypothyroidism increases the risk of statin-induced adverse muscle effects.

Nephrotic syndrome

Longstanding nephrotic syndrome is associated with accelerated atherosclerosis; thus lipid-lowering treatment will usually be required. There is evidence that, in the nephrotic syndrome, the hyperlipidaemia may worsen renal function: oxidized LDL has been shown in vitro to affect mesangial cells and accelerate glomerulosclerosis. The hyperlipidaemia occurring in the nephrotic syndrome is most commonly hypercholesterolaemia, but mixed hyperlipidaemia also occurs. The hypercholesterolaemia is a result of hepatic overproduction of apo B-100 as part of the generally increased hepatic protein synthesis typical of the condition. Hydroxymethylglutaryl-CoA (HMG-CoA) reductase activity is also increased. The increase in triglyceride results from reduced removal of chylomicrons and VLDL due to impaired lipoprotein lipase activity. HDL may be lost in the urine and, dependent on whether or not increased synthesis matches the rate of loss, the plasma HDL-cholesterol concentration may be low, normal or occasionally high.

Chronic kidney disease

Patients with chronic kidney disease (CKD) are at very high risk of cardiovascular disease. The lipid abnormality most commonly seen is hypertriglyceridaemia; the lipoprotein profile is characterized by a high concentration of intestinally derived apo B-48-containing lipoproteins and their remnants. The composition of the lipoproteins is also changed. The apo C-III concentration of the triglyceride-rich lipoproteins is increased and, since apo C-III inhibits both lipoprotein lipase and the hepatic uptake of chylomicron and VLDL remnants, this results in

increased plasma triglyceride concentrations. Although total cholesterol may be relatively 'normal', much of the LDL is small and dense and therefore is particularly atherogenic. Although hypertriglyceridaemia is usually present, the most extensive evidence for reduction in cardiovascular disease risk by lipid modification is through use of HMGCoA reductase inhibitors (statins), which reduce LDL, even though they have comparatively little effect on plasma triglycerides.

Renal transplantation

After transplantation, the hyperlipidaemia that accompanies impaired renal function may be corrected, but this is not always the case. In addition, immunosuppressant therapy, including corticosteroids, may itself cause hyperlipidaemia. The degree of hypercholesterolaemia is often greater with sirolimus and ciclosporin than with tacrolimus. Thus, subjects post-transplant remain at high risk of cardiovascular disease, and since they will have been hyperlipidaemic prior to transplantation, therapy to improve their lipids is warranted. Caution needs to be taken, however, because of the well-documented interactions between ciclosporin and both statins and fibrates.

Liver disease

The liver plays a central role in lipoprotein metabolism. In addition, the only physiological means the body has of excreting cholesterol is by its hepatic secretion into the bile.

Cholestasis is frequently accompanied by a mixed hyperlipidaemia due to the accumulation of remnant lipoproteins. Lipoprotein X is found uniquely in cholestasis. This lipoprotein contains bile acids, apo C, apo D, albumin and cholesterol, and while it falls within the LDL density range, unlike all the other lipoproteins, it migrates towards the cathode on lipoprotein electrophoresis. The atherogenic potential of lipoprotein X is undefined; when present in massive amounts (cholesterol >50mmol/L), lipoprotein X has been associated with hyperviscosity. Plasma Lp(a) concentration is low and HDL-cholesterol is high in cholestasis; these abnormalities may, in part, explain the fact that cholestasis is not generally associated with a high cardiovascular risk.

Non-alcoholic fatty liver disease (NAFLD) is a manifestation of the metabolic syndrome and, as such, is associated with an increased cardiovascular risk. The dyslipidaemia associated with NAFLD is either hypertriglyceridaemia or mixed hyperlipidaemia; it is frequently associated with small, dense LDL particles. Where the condition runs in families, it is difficult to distinguish from familial combined hyperlipidaemia. Subjects with NAFLD may have abnormal plasma liver enzyme activities (particularly increased aminotransferases) and caution therefore needs to be exercised when starting lipid-lowering agents. However, owing to the increased cardiovascular risk, treatment is warranted. Statins are generally the drug of first choice unless severe hypertriglyceridaemia is present creating a risk of pancreatitis. Although liver enzymes should be monitored, treatment is usually safe, and in some individuals, normalization of the lipid profile results in a reduction in the aminotransferase activities.

Alcohol

Alcohol causes hypertriglyceridaemia in susceptible individuals. This results from a combination of increased production and impaired removal of VLDL. In severe cases, this may result in chylomicronaemia, which may result in acute pancreatitis.

Epidemiological studies have suggested that moderate alcohol intake (no more than 1–2 units per day) is associated with a lower mortality than either higher alcohol intakes or abstinence from alcohol. Such moderate intake is associated with an increase in plasma HDL-cholesterol concentration, which may be responsible for the apparent cardiovascular protection.

Drug-related hyperlipidaemia

Various drugs, including antihypertensive agents (β -blockers, thiazides), corticosteroids, sex steroids, immunosuppressants, second generation antipsychotics and antiretroviral drugs, can affect lipoprotein concentrations.

The effect of antihypertensives on the lipid profile has been extensively studied since early hypertension trials demonstrated that effective lowering of blood pressure reduced the incidence of strokes but not that of coronary heart disease. The effect of β -blockers depends on their selectivity: non-selective (e.g. propranolol) and β_1 -selective (e.g. atenolol) β -blockers raise plasma triglycerides and lower HDL, whereas β -blockers with intrinsic sympathomimetic activity (e.g. pindolol) are lipid neutral. The hyperlipidaemia associated with diuretic therapy appears to be caused by the unopposed α -adrenergic activity elicited by these drugs; the low dosages (e.g. 2.5 mg of bendroflumethiazide) now used for treatment of hypertension have a negligible effect on the lipid profile.

Glucocorticoid treatment results in an increase in plasma LDL, HDL and triglyceride concentrations. Oestrogens increase hepatic VLDL secretion and increase HDL (HDL₂). In postmenopausal women, they also tend to reduce LDL. Progestogens, on the other hand, tend to reduce HDL (especially HDL₂) and raise LDL. With both oral contraceptives and hormone replacement therapy, the magnitude of the changes in the lipoproteins depends on the dose, the route of administration, the particular oestrogen and progestogen involved and whether or not there is an underlying primary hyperlipoproteinaemia. The use of oral oestrogen results in exposure of the liver to supra-physiological doses of oestrogen, with stimulation of hepatic lipoprotein synthesis. Severe hypertriglyceridaemia may be precipitated by the use of oral oestrogen in susceptible individuals, who may have a normal plasma triglyceride concentration or only mild hypertriglyceridaemia when not taking oestrogen or using topical (patch) oestrogen hormone replacement therapy. Of the progestogens, the 19-nortestosterone derivatives (e.g. norethisterone, levonorgestrel) are considerably more androgenic than the C21 progestogens (e.g. medroxyprogesterone acetate, dydrogesterone) and result in greater reductions in plasma HDL concentrations.

Tamoxifen, a selective oestrogen receptor modulator that has agonist and antagonist oestrogen effects, has a predominantly agonistic effect on the liver and thus can cause severe hypertriglyceridaemia.

Among immunosuppressant drugs, both the calcineurin inhibitors (ciclosporin, tacrolimus) and the non-calcineurin inhibitor, sirolimus, can cause hyperlipidaemia, which, in most individuals, is an isolated increase in plasma LDL concentration.

Hypertriglyceridaemia is a common side-effect of second generation antipsychotic drugs including clozapine, olanzapine and quetiapine. This may be severe, and cases of pancreatitis, thought to result from the hypertriglyceridaemia, have been reported.

Antiretroviral drug treatment is frequently associated with dyslipidaemia. The combination of fat redistribution, insulin resistance and dyslipidaemia seen in subjects on antiretroviral treatment is referred to as 'the lipodystrophy syndrome'. The fat redistribution is characterized by loss of subcutaneous fat and accumulation of intra-abdominal fat. This is associated with regimens containing protease inhibitors and nucleoside reverse transcriptase inhibitors, whereas the dyslipidaemia and insulin resistance are associated particularly with protease inhibitors. Care is required in treating dyslipidaemia in these individuals as a number of antiretroviral therapies interact with lipid lowering treatments.

ACQUIRED HYPOLIPIDAEMIA

Acquired or secondary hypolipidaemia is much less common than its hyperlipidaemic equivalent. It occurs with cachexia, malabsorption, hyperthyroidism, malnutrition, liver failure and some forms of malignancy. It is of no direct clinical consequence, being only a manifestation of the underlying condition.

Drug therapy can also cause hypolipidaemia. Excessive thyroxine will cause a reduction in total cholesterol, mimicking that seen with thyrotoxicosis. Drugs including fibrates and thiazolidinediones have been reported occasionally to cause an unpredictable and profound paradoxical fall in plasma HDL-cholesterol concentration in susceptible individuals.

INVESTIGATION OF LIPID DISORDERS

The simple appearance of a serum or plasma sample may indicate a lipid disorder. Chylomicrons and VLDL are large enough to scatter light. Chylomicrons are less dense than plasma and will form a layer on the surface of a sample left standing at 4°C overnight. If present in very large amounts, VLDL will make the sample appear opalescent.

Total cholesterol

This is the most commonly used single measure of lipid status. It has been used extensively in epidemiological studies, and may have some use in monitoring treatment. However, it is inadequate on its own for the diagnosis of lipid disorders, or as the only measure prior to starting

treatment. A full fasting lipid profile including triglyceride and HDL-cholesterol should be measured at least once to avoid missing significant dyslipidaemias with normal total cholesterol. LDL-cholesterol, non-HDL-cholesterol and apo B measurements may have benefits in assessing cardiovascular risk and the adequacy of treatment.

Cholesterol is generally measured by enzymatic methods using cholesterol oxidase. Although analytical CV is low (typically <3%), intraindividual biological variation is around 5% so it is important to appreciate that the critical difference between consecutive measurements in monitoring therapy may be around 0.8 mmol/L. Fasting or non-fasting makes little difference ($\pm 3\%$) to measurements.

Triglycerides

Triglyceride increases up to 2–3-fold after a meal, and so samples should be taken after an overnight (>12 h) fast in order to obviate difficulties of interpretation resulting from the presence of chylomicrons or chylomicron remnants. Most triglyceride methods involve hydrolysis of triglycerides and measurement of the free glycerol released. Glycerol blanking is not routinely performed. In certain circumstances, such as patients with uncontrolled diabetes mellitus, on haemodialysis, or with the rare X-linked recessive disorder glycerol kinase deficiency, the glycerol content of the sample may be significantly raised. In these situations, high concentrations of glycerol in the plasma will produce falsely high triglyceride results, and glycerol blanking is essential to obtain a valid result.

Plasma triglyceride concentration has a much greater biological variation than does that of cholesterol, at ~20% even in fasting samples.

High density lipoprotein cholesterol

Plasma HDL concentration is generally measured by direct (homogeneous) methods that rely on antigen/antibody complex formation, or polyethylene glycol (PEG) modified enzymes that react selectively with cholesterol in HDL. These assays can provide reliable results in the presence of triglyceride concentrations up to 10 mmol/L.

Low density lipoprotein cholesterol

LDL-cholesterol (LDL-C) can be derived by substituting the results of the analysis of total cholesterol (TC), HDL-cholesterol (HDL-C) and triglyceride (TG) (fasting) into the Friedewald formula:

$$\text{LDL-C} = \text{TC} - \left(\text{HDL-C} + \frac{\text{TG}}{2.2} \right)$$

(all measurements being in mmol/L).

The Friedewald formula assumes that most of the triglyceride in the plasma is in VLDL and that there is a molar ratio of 5:1 of triglyceride to cholesterol in the VLDL fraction. For this reason, the Friedewald formula is not applicable in subjects with remnant hyperlipidaemia in whom the remnant lipoprotein fraction is cholesterol-enriched. In addition, it is not applicable to subjects with plasma

triglyceride concentrations >4.5 mmol/L since, at these levels, VLDL contains a greater proportion of triglyceride, and so the formula overestimates VLDL-cholesterol and underestimates LDL-cholesterol. Although most laboratories report derived plasma LDL-cholesterol concentrations on samples with triglycerides up to 4.5 mmol/L this tendency for VLDL to contain a greater proportion of triglyceride is sometimes present at lower concentrations of triglycerides, to a degree which may be clinically significant. The error in calculated LDL-cholesterol is $>10\%$ in more than 30% of subjects with a plasma triglyceride concentration of 2.3–3.4 mmol/L and in $>40\%$ of subjects with triglycerides of 3.4–4.5 mmol/L.

Directly LDL-cholesterol assays are being used increasingly. In general, they compare well with LDL-cholesterol as measured using ultracentrifugation. They have the advantage of being able to measure LDL-cholesterol in non-fasting samples, but because they are immunoseparation methods, have the disadvantage of higher cost compared with the simpler enzymatic measurement of total and HDL-cholesterol, and triglyceride.

Plasma LDL-cholesterol concentration is increasingly used in epidemiological and therapeutic studies and treatment goals based on LDL-cholesterol have been incorporated into most national guidelines.

Non-HDL-cholesterol

Non-HDL-cholesterol is derived simply using the formula (total cholesterol – HDL-cholesterol). The major difference from LDL-cholesterol is that it includes cholesterol in VLDL. In some epidemiological studies it has been shown to predict vascular risk almost as well as apo B and better than LDL-cholesterol. This may simply be because VLDL is atherogenic in its own right, but also because non-HDL-cholesterol may reflect LDL particle number better than measurement of LDL-cholesterol concentration alone.

Other advantages of non-HDL-cholesterol are that it can be measured in non-fasting samples, and that its measurement has better precision and accuracy than Friedewald derived LDL-cholesterol. It is also cheaper than measurement of apo B. It has been adopted in a number of current guidelines alongside LDL-cholesterol.

Apolipoproteins

Apolipoprotein A-I

Apolipoprotein A-I measurement has been advocated as an alternative to HDL-cholesterol, but it is generally more expensive, and clear advantage over measurement of HDL-cholesterol with modern homogeneous assays has not been demonstrated.

Apolipoprotein B

Apolipoprotein B measurement does serve a useful purpose in the investigation of lipid disorders. As there is only one apo B molecule in each apo B containing lipoprotein particle, measurement of apo B provides a measure of particle number. There is epidemiological evidence that

this may provide a better predictor of cardiovascular risk and response to treatment than measurement of LDL-cholesterol (directly measured or derived). This is because small dense LDL particles will tend to result in lower concentrations of LDL-cholesterol, but are relatively more atherogenic than less dense, cholesterol-enriched LDL.

Apolipoprotein B measurement may be helpful in diagnosis, particularly in patients with mixed hyperlipidaemia. Apo B is often elevated in familial combined hyperlipidaemia, and low in proportion to total cholesterol in remnant hyperlipidaemia.

Apolipoprotein B and non-HDL-cholesterol show similar ability to predict cardiac risk in population studies. Apolipoprotein B may, however, be superior in patients with a large proportion of small dense LDL including those with type 2 diabetes and the metabolic syndrome.

Apolipoprotein E

Although it is not essential for managing patients, knowledge of the apolipoprotein E phenotype (based on electrophoresis) or confirmation of an $\epsilon 2/\epsilon 2$ genotype is sometimes helpful in confirming a diagnosis of remnant hyperlipoproteinaemia. It should also be remembered that the finding of an $\epsilon 2/\epsilon 2$ genotype on its own does not establish the diagnosis as other factors need to be present before the typical clinical phenotype becomes manifest.

Lipoprotein(a)

Standardization has only become available for Lp(a) over the last few years. Very wide variation between methods made the assay of limited utility in the past. It is important that assays are independent of the number of kringle 4 repeats in the highly variable region of the molecule. Plasma Lp(a) concentration is normally <30 mg/dL but shows a highly positively skewed distribution. High concentrations are associated with an increased risk of cardiovascular disease, and the greatest utility of measurement is in assessing risk in those with high plasma LDL-cholesterol concentration.

In patients with familial hypercholesterolaemia who have very high Lp(a) concentrations (>60 mg/dL), LDL apheresis, which selectively removes apo B containing lipoprotein particles, is occasionally used, as statins have very little impact on Lp(a).

Post-heparin lipolytic activity

The intravenous injection of heparin leads to the release into the circulation of both LPL and HL. Measurement of LPL activity is useful in the differential diagnosis of severe hypertriglyceridaemia where chylomicronaemia is present (type I hyperlipidaemia). This may be due to a deficiency of lipoprotein lipase.

Lipoprotein separation techniques

Ultracentrifugation

Preparative ultracentrifugation is the reference procedure for isolating lipoproteins; it involves sequential isolation of lipoprotein classes from plasma after adjustment

of the density with sodium chloride, potassium bromide, sodium bromide or mixtures of these salts. This is a time-consuming procedure and may take up to five days for isolation of all subclasses to be completed. Losses may occur at all stages and artefacts may also be introduced. It is generally reserved for research studies.

Lipoprotein electrophoresis

Electrophoresis on cellulose acetate, paper, agarose and polyacrylamide has been used as a means of separating the lipoprotein classes and, indeed, the separation pattern on paper electrophoresis was the basis of the Fredrickson classification of lipid disorders. For therapeutic purposes, the Fredrickson classification has been replaced by defining the genetic basis of the hyperlipidaemia. It may still be of use in the diagnosis of remnant hyperlipidaemia with its broad β band, and may be helpful in deficiency states; for example, in demonstrating the lack of pre- β and β bands in abetalipoproteinaemia.

Genotyping

Genetic testing is particularly used for the diagnosis of remnant hyperlipidaemia, and familial hypercholesterolaemia (FH). Homozygosity for the $\epsilon 2$ allele of the apolipoprotein E gene is seen in remnant hyperlipidaemia. FH is due to mutations in *LDLR*, *APOB*, *PCSK9* or *LDLRAP1* genes.

Genotyping will also be used increasingly to confirm the diagnosis of rare monogenic lipid disorders. In addition, genome-wide association studies are starting to identify single nucleotide polymorphisms (SNPs) in multiple genes that may have minor effects individually but clinically significant effects cumulatively. This is starting to provide insight into the genetic basis of many common lipid and lipoprotein phenotypes, but has not yet found application in routine practice.

A further application of genetic testing is in tailoring therapy. A SNP in *SLCO1B1* has been described which is associated with a four-fold increased risk increased risk of statin toxicity in heterozygotes and 16-fold increased risk in homozygotes. This gene codes for an organic anion transporter protein (OATP1B1) that is responsible for the uptake of statins to their site of action in the liver. Deficiency results in reduced efficacy and increased plasma levels with consequent increased risk of myositis. Fluvastatin is not taken up via this mechanism, and may be more effective and better tolerated in individuals with this polymorphism.

TREATMENT OF HYPERLIPIDAEMIA

Management of patients with hyperlipidaemia requires a full history, including a personal history of cardiovascular disease and conditions that can cause dyslipidaemias, family history of cardiovascular disease (with age of onset), lifestyle factors and drugs that may produce lipid abnormalities. Clinical examination and a detailed family history may suggest a specific genetic diagnosis, which may be confirmed by lipid, lipoprotein or genetic testing.

Genetic lipid disorders may require more aggressive treatment, particularly if they suggest (e.g. in FH) that cholesterol levels have been elevated since soon after birth. They may also indicate the need for family screening.

All patients should receive lifestyle advice (including advice concerning smoking cessation where relevant). The aim of dietary advice is both to achieve weight loss (where needed) and an appropriate balance of nutrients including control of the amount of fat, and especially saturated fat, in the diet. Causes of secondary dyslipidaemia should be treated: correction of untreated hypothyroidism or newly diagnosed diabetes may produce a marked fall in plasma lipid concentrations.

In patients with fasting serum triglyceride concentrations of >10 mmol/L, there is a risk of acute pancreatitis, and treatment to lower the triglycerides should normally be the initial priority. Fibric acid derivatives (fibrates) and high doses of fish oils (e.g. Omacor®) are generally the first-line treatments. Some patients with type 2 diabetes may benefit from the introduction of insulin.

In other cases, treatment is based on serum lipid concentrations and an assessment of a patient's overall cardiovascular risk. Reduction in cardiovascular risk is the main aim, and statins are generally the first-line treatment. They have the most extensive evidence base and are generally used even if triglyceride concentrations are slightly or moderately raised, so long as this is not high enough to cause pancreatitis.

The highest priority is to treat those with clinical evidence of vascular disease (including cerebrovascular and peripheral vascular as well as coronary disease, i.e. secondary prevention). Primary prevention (i.e. in the absence of vascular disease) is targeted to those at the highest risk. This includes those with genetic hyperlipidaemias (e.g. FH), those with conditions known to cause qualitative lipoprotein changes that increase risk (e.g. diabetes, when treatment is generally recommended above the age of 40), and those estimated to be at high risk of cardiovascular disease based on lipid results and cardiovascular risk factors. Various risk calculation tools are available for calculation of ten-year or lifetime risk, based on epidemiological data (see Chapter 38 for a more detailed discussion).

The aim of treatment is ideally to reduce serum total cholesterol concentration to <4 mmol/L and LDL-cholesterol concentration to <2 mmol/L. Targets for non-HDL-cholesterol of <2.5 mmol/L and apo B of <0.8 g/L are also recommended. These targets recognize that residual risk of cardiovascular event after LDL-cholesterol reduction with statins depends on LDL particle number, which may better be reflected by measuring plasma apo B concentration.

If targets are not achieved, then fibrates, ezetimibe or bile acid sequestrants are all options, usually in addition to a statin, although none have such a robust evidence base as statins for reducing cardiovascular events. Use of lipid lowering drugs in combination is usually safe although care must be taken to identify patients who are at increased risk of skeletal myopathy.

In patients with FH, drug treatment is generally considered by the age of ten, with the aim of reducing serum LDL-cholesterol concentration by $>50\%$. In those with a more severe phenotype or with homozygous FH, LDL

TABLE 37.7 Drugs used in the treatment of dyslipoproteinaemias

Drug group	Mode of action	Indications/advantages	Adverse effects
Bile acid sequestrants (e.g. cholestyramine, colestipol, colesevelam)	Interrupt the enterohepatic circulation by binding bile salts. This results in increased hepatic LDL uptake but also increased cholesterol synthesis	Raised LDL-C. Safe in pregnancy. Used in LRCCT ^a and shown to reduce morbidity and mortality from coronary artery disease	Increase triglyceride concentration. Constipation, diarrhoea, nausea, abdominal bloating or discomfort. May interfere with absorption of other drugs
Fibric acid derivatives (e.g. fenofibrate, bezafibrate, gemfibrozil)	PPAR α agonist activity has multiple effects including increased LPL activity, increased <i>ABCA1</i> transcription, reduced hepatic VLDL production, increased LDLR mediated uptake and increased biliary cholesterol secretion	Hypertriglyceridaemia (\uparrow VLDL) Mixed hyperlipidaemia (\uparrow VLDL and \uparrow LDL) and remnant hyperlipoproteinaemia. Tendency for HDL-C to increase	Increased creatine kinase and myopathy – care necessary if using in combination with a statin (avoid gemfibrozil). Indigestion. Cholelithiasis. Potentiation of oral anticoagulants. Abnormal liver enzymes
Hydroxymethoxyglutaryl-coenzyme A reductase inhibitors (e.g. simvastatin, pravastatin, fluvastatin, atorvastatin and rosuvastatin)	Inhibit the rate-limiting enzyme of cholesterol synthesis thus increasing LDL receptor-mediated uptake from plasma	Raised LDL-C. Well tolerated. Clinical endpoint studies showing benefits in reducing cardiovascular events and total mortality (CTT ^b). ASTEROID ^c (rosuvastatin) showed regression of coronary atherosclerosis	Occasional patients develop myopathy and rhabdomyolysis, especially with concomitant use of drugs including ciclosporin, gemfibrozil and erythromycin. Abnormal liver enzymes
ω -3 Fish oil rich in the polyunsaturated fatty acids: eicosapentaenoic and docosahexanoic acids, e.g. Omacor [®] , Maxepa [®]	Reduces VLDL triglyceride synthesis	High triglyceride concentrations due to increased VLDL, whether alone or in combination with chylomicronaemia. At pharmacological dosage also possesses antithrombotic and anti-inflammatory properties	Nausea, flatulence
Ezetimibe	Specific inhibitor of intestinal cholesterol absorption. Blocks enterohepatic recirculation of biliary cholesterol, as well as reducing absorption of dietary cholesterol	As an adjunct to a statin for the treatment of elevated LDL-C or alone if a statin is not tolerated or inappropriate. Can also be used to treat β -sitosterolaemia (phytosterolaemia)	Gastrointestinal disturbance. Raised creatine kinase or myalgia
Drugs in clinical trials			
CETP inhibitors (anacetrapib; evacetrapib)	Inhibit cholesterol ester transfer protein	As adjunct to statins to raise HDL-C and reduce cardiovascular risk	Effect of increasing HDL-C on cardiovascular outcome still uncertain.
MTP inhibitors (lomitapide)	Inhibit microsomal triglyceride transfer protein	Clinical trials ongoing in homozygous FH and in familial hypertriglyceridaemia	Hepatic steatosis
LDL antisense oligonucleotides, e.g. mipomersen (KYNAMRO [®])	Reduce LDL synthesis by inhibiting <i>APOB</i> transcription	Hypercholesterolaemia	
PCSK9 inhibitors	Increase LDL receptor activity by genetic inhibition or immunological blockade of PCSK9	Hypercholesterolaemia	Reduce HDL-C as well as LDL-C cholesterol. Injectable. Effect on outcome unclear
Drugs under development			
Apo A-I based therapies	Apo A-I mimetics and synthetic HDL particle infusion		
LPL gene replacement ABCA1 agonists LCAT activators Lp-PLA ₂ inhibitors		LPL deficiency	

^aLRCCT, Lipid Research Clinics Coronary Primary Prevention Trial^bCTT, Cholesterol Treatment Trialists collaboration (a meta-analysis of the major statin trials)^cASTEROID, A study to evaluate the effect of rosuvastatin on intravascular ultrasound derived coronary atherosclerosis burden.

apheresis is a treatment option. New agents such as MTP and PCSK9 inhibitors, and apo B mRNA antisense oligonucleotides are in development or clinical trials.

A summary of drugs used to treat hyperlipidaemia, and of those currently in development or clinical trials, is given in [Table 37.7](#).

Side-effects of statins and other lipid lowering treatments include elevation of transaminases and muscle effects ranging from an asymptomatic slight rise in creatine kinase (CK), through myalgia or myositis, with or without a rise in CK, to, rarely, rhabdomyolysis. Patients should be warned of possible effects and liver enzymes should be monitored. Drug interactions (e.g. with immunosuppressant drugs) increase the risk of myositis. Genetic variants may also increase the risk of muscle side-effects (see p. 734).

While targeting those at highest risk, and treating them aggressively with lipid lowering drug treatment, is highly effective in individuals, it is important to recognize that the greatest impact on population rates of CVD would still to be made by changing the diet and lifestyle habits of the population as a whole.

CONCLUSION

The principal lipids present in the blood are cholesterol, triglycerides and phospholipids. They are transported, in association with various apolipoproteins, in lipoprotein particles. Lipids have essential functions in cellular structure and metabolism but, at present, the major clinical interest in them derives from the relationship between the plasma cholesterol concentration (specifically LDL-cholesterol) and the risk of developing atherosclerosis.

Lipid metabolism is complex; there is a continuous flux of lipoprotein particles between the blood and tissues, and of the lipid and protein components between lipoprotein particles. Abnormalities of plasma lipids and of

lipoprotein composition occur frequently. They can arise as a result of a disease that secondarily affects lipid metabolism (e.g. diabetes, hypothyroidism) or be primary, that is, genetically determined.

In the genetic hyperlipidaemias, appropriate investigation will define the nature and severity of the disorder and provide a rational basis for treatment. Greater understanding of the details of lipoprotein metabolism will continue to identify new therapeutic targets and bring the potential for effective new therapies. In the majority of cases, the aim is to reduce the risk of cardiovascular disease, and therefore specific therapy for the hyperlipidaemia needs to be accompanied by action against all the other pertinent risk factors.

ACKNOWLEDGEMENT

The author wishes to acknowledge Dr Christine Marenah, who was the author of this chapter in previous editions of the book.

Further reading

Cholesterol Treatment Trialists Collaboration. Efficacy and safety of more intensive lowering of LDL cholesterol: a meta analysis of data from 170,000 participants in 26 randomised trials. *Lancet* 2010;376:1670–81.

This meta analysis reviews the major trials of statin therapy and their effect on cardiovascular outcomes.

Dennis EA, Witztum JL, editors-in-chief. *Journal of Lipid Research*.

This monthly publication includes reviews on a wide range of lipid-related topics.

Grundy SM, editor-in-chief. *Current Opinion in Lipidology*.

This two-monthly journal includes articles and reviews of recent literature.

Rifai N, Warnick GR, Dominiczak MH, editors. *Handbook of lipoprotein testing*. 2nd ed Washington: AACCC Press; 2000.

A comprehensive textbook of analytical procedures in relation to lipid metabolism.

Scriver CR, Beaudet AL, Sly WS et al. editors. *The metabolic and molecular bases of inherited disease*. 8th ed New York: McGraw-Hill; 2001. p. 2705–960, Part 12 (Lipids).

This is the most recent hard copy version of this textbook. Updated sections are available online at www.ommbid.com;

Clinical biochemistry of the cardiovascular system

Clodagh M. Loughrey • Ian S. Young

CHAPTER OUTLINE

INTRODUCTION 737

Cardiovascular disease 737

Role of the laboratory 739

CARDIAC MUSCLE STRUCTURE AND BIOCHEMISTRY 740

ARTERIAL STRUCTURE AND FUNCTION 741

ATHEROSCLEROSIS 741

ACUTE MYOCARDIAL DAMAGE 744

Biomarkers of acute myocardial
damage 744

HEART FAILURE 749

Natriuretic peptides 749

CARDIOVASCULAR RISK FACTORS 750

Cardiovascular risk assessment 750

Unmodifiable risk factors 751

Potentially modifiable risk factors 752

Dietary factors 756

HYPERTENSION 757

Definition 757

Cause 758

Laboratory assessment of
hypertension 759

Renovascular hypertension 759

Primary aldosteronism
(hyperaldosteronism) 760

Phaeochromocytoma 763

Malignant hypertension 764

Hypertension in pregnancy 764

Management of hypertension 764

CONCLUSION 765

APPENDIX 765

INTRODUCTION

The cardiovascular or circulatory system consists of the heart and blood vessels (arteries, arterioles, capillaries, venules and veins). When it functions normally, effective circulation of blood maintains perfusion of the tissues, so that substrates for cellular metabolism are provided and excretory products removed. Among other vital functions, it also allows hormones to be transported from their organs of origin to their target tissues, defends against infection through facilitation of the movement of white cells and cytokines, and promotes haemostasis through delivery of platelets and clotting factors to traumatized tissue.

Circulation of blood is maintained by the pumping action of the heart, a muscular organ consisting of two atria and two ventricles. The atria receive blood (the left atrium from the lungs and the right atrium from the rest of the body) and pass it to the ventricles. The right ventricle pumps deoxygenated venous blood to the lungs for oxygenation while the left ventricle supplies oxygenated blood to the rest of the body (including itself) via the aorta, arteries, arterioles and capillaries.

Cardiovascular disease

Cardiovascular disease (CVD) collectively comprises disease of the heart and of blood vessels (almost invariably arteries) supplying any organ. The organs most commonly affected by arterial disease are the heart (coronary artery disease), the brain (cerebrovascular disease) and the limbs (peripheral arterial disease). Renovascular disease is an important cause of chronic kidney disease and of hypertension.

Cardiovascular disease is the leading cause of death in England and Wales, currently responsible for one in three deaths. For every death due to CVD, there are at least two major non-fatal CVD events. It is also the leading cause of death globally: the World Health Organization (WHO) estimated that 17.3 million people died from CVD in 2008, representing 30% of all deaths. Of these, an estimated 7.3 million deaths were due to coronary heart disease and 6.2 million were due to stroke (cerebrovascular disease). Although approximately 80% of CVD is due to modifiable factors and is potentially preventable, CVD deaths continue to rise, largely because preventative measures are insufficient or ineffective. Low- and middle-income countries are disproportionately affected,

with less than 20% of global deaths from CVD occurring in high-income countries (Fig. 38.1). Mortality due to all main communicable diseases, including HIV/AIDS, tuberculosis and malaria, is expected to decline globally between now and 2030. By 2030, it is anticipated that CVD alone will be responsible for more deaths in low-income countries than infectious diseases, maternal and perinatal conditions, and nutritional disorders combined, driven in particular by the increased global prevalence of obesity. Worldwide, it is estimated that in 2030, non-communicable diseases will account for more than three-quarters of all deaths, and almost 23.6 million people will die from CVD. Thus, CVD is today the largest single contributor to global mortality and is set to continue to dominate mortality trends on the world stage well into the future.

Atherosclerosis is by far the most common cause of cardiovascular disease in developed countries: thus the term ‘cardiovascular disease’ is generally used and perceived to mean ‘atherosclerotic disease’, the term ‘coronary artery disease’ or ‘coronary heart disease’ to mean atherosclerosis

of the coronary arteries, and ‘cerebrovascular disease’ to mean atherosclerosis of the cerebral vasculature. Coronary artery atherosclerosis leads to diminished blood supply to, or ischaemia of, the myocardium. When this becomes critical the heart muscles die, leading to a ‘myocardial infarction’ (MI), often, but not invariably, accompanied by diagnostic changes on the electrocardiogram (ECG). Acute coronary syndrome (ACS) is a blanket term for patients presenting acutely with clinical features of myocardial ischaemia, and encompasses ST elevation MI (STEMI), non-ST elevation MI (NSTEMI) and unstable angina (UA) (Fig. 38.2). Death from ischaemic heart disease leading to failure of the heart to maintain circulation may occur acutely in the setting of a myocardial infarction (either due to a large area of infarction or to cardiac dysrhythmia, which interferes with effective ventricular contraction) or in the setting of chronic heart failure. Treatment of MI with thrombolytic drugs or percutaneous coronary intervention (PCI) with or without stenting can be life-saving but depends on prompt and accurate diagnosis.

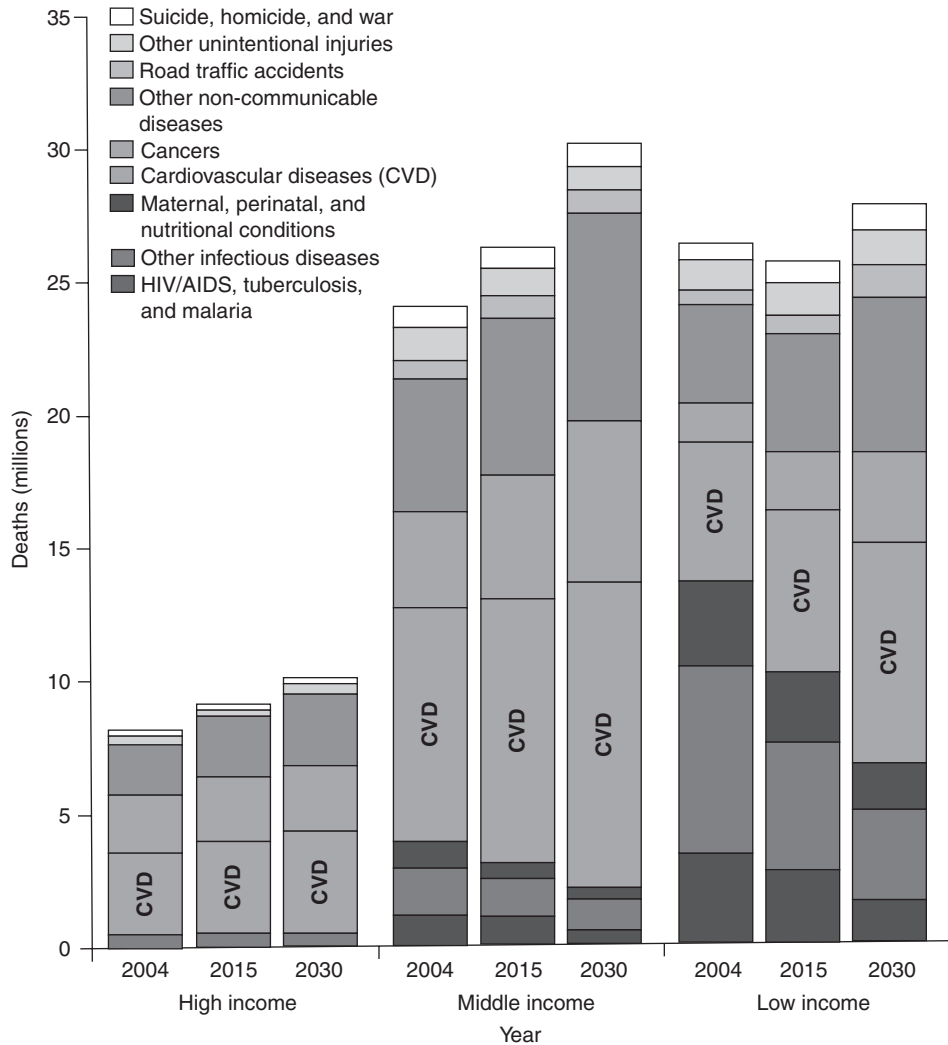


FIGURE 38.1 ■ Projected deaths by cause in 2004, 2015 and 2030, in low-, middle- and high-income countries. Countries are categorized by the World Bank according to 2004 gross national income per person: low income < US\$826; middle income US\$826–\$10 065; high income > US\$10 066. Adapted from Figure 2 in Beaglehole R, Bonita R 2008 Global health: a scorecard. *Lancet* 372: 1988–1996, with permission.

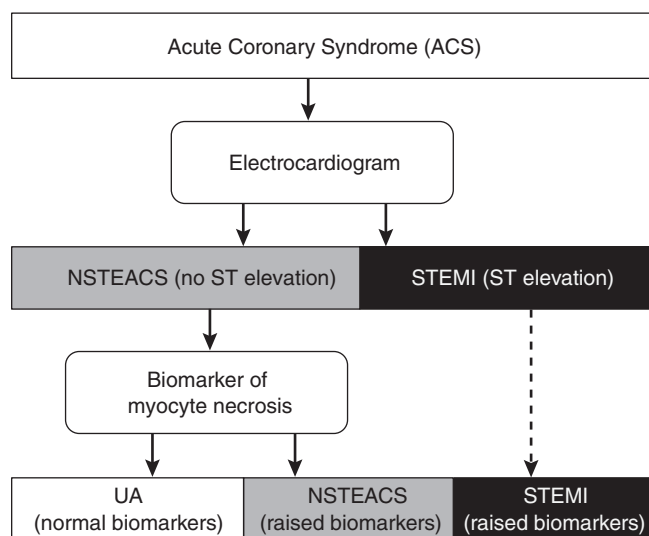


FIGURE 38.2 ■ Categories of acute coronary syndromes, and the investigations used to discriminate between them. NSTEMACS, non-ST elevation ACS; STEMI, ST elevation myocardial infarction; UA, unstable angina.

Role of the laboratory

The clinical biochemistry laboratory plays a critical role in the management of acute myocardial ischaemia, where delay in diagnosis and treatment directly increases morbidity and mortality. For many years, the diagnosis of acute MI relied on criteria that had been established by the World Health Organization (WHO) in 1979: rise in a biomarker of cardiac myocyte injury was one of the criteria but was not a prerequisite. The development of more specific and sensitive cardiac biomarkers has facilitated earlier and more accurate diagnosis, which permits more timely and appropriate intervention. In 2000, recognition of the fact that the newer cardiac biomarkers could identify myocardial micronecrosis, even when conventional diagnostic criteria excluded MI, led to a new joint European Society of Cardiology and American College of Cardiology (ESC/ACC) committee universal definition of 'myocardial infarction', in which biomarker measurement played a fundamental role. The definition was further revised in 2007 and again in 2012 by the Joint Task Force of the European Society of Cardiology, American College of Cardiology Foundation, the American Heart Association and the World Health Federation (ESC/ACCF/AHA/WHF). The current universal definition of acute MI has a rise and/or fall in cardiac biomarkers (preferably troponin) as the central feature of the definition (Box 38.1).

Identification of those at high risk of developing coronary atherosclerosis allows earlier intervention to prevent related clinical events, and there has also been sustained focus on developing biomarkers to aid prediction of cardiovascular risk. The laboratory also plays important roles in the diagnosis of heart failure and in the investigation of hypertension.

After sections describing the structure of heart muscle and arteries, and the pathogenesis of atherosclerosis, these roles will be explored in turn.

BOX 38.1 Definition of myocardial infarction

Criteria for acute myocardial infarction

The term acute myocardial infarction (MI) should be used when there is evidence of myocardial necrosis in a clinical setting consistent with acute myocardial ischaemia.

Under these conditions any one of the following criteria meets the diagnosis for MI.

- Detection of a rise and/or fall of cardiac biomarker values (preferably cardiac troponin, cTn) with at least one value above the 99th percentile upper reference limit (URL) and with at least one of the following:
 - symptoms of ischaemia
 - new or presumed new significant ST-segment-T wave (ST-T) changes or new left bundle branch block (LBBB)
 - development of pathological Q waves in the ECG
 - imaging evidence of new loss of viable myocardium or new regional wall motion abnormality
 - identification of an intracoronary thrombus by angiography or autopsy.
- Cardiac death with symptoms suggestive of myocardial ischaemia and presumed new ischaemic ECG changes or new LBBB, but death occurred before cardiac biomarkers were obtained or before cardiac biomarker values would be increased.
- Percutaneous coronary intervention (PCI) related MI is arbitrarily defined by elevation of cTn values ($>5 \times 99$ th centile URL) in patients with normal baseline values (≤ 99 th centile URL) or a rise of cTn values $>20\%$ if the baseline values are elevated and are stable or falling. In addition, either: (i) symptoms suggestive of myocardial ischaemia or (ii) new ischaemic ECG changes or (iii) angiographic findings consistent with a procedural complication or (iv) imaging demonstration of new loss of viable myocardium or new regional wall motion abnormality are required.
- Stent thrombosis associated with MI when detected by coronary angiography or autopsy in the setting of myocardial ischaemia and with a rise and/or fall of cardiac biomarker values with at least one value above the 99th percentile URL.
- Coronary artery bypass grafting (CABG)-related MI is arbitrarily defined by elevation of cardiac biomarker values ($>10 \times 99$ th centile URL) in patients with normal baseline cTn values (≤ 99 th centile URL). In addition, either (i) new pathological Q waves or new LBBB or (ii) angiographic documented new graft or new native coronary artery occlusion or (iii) imaging evidence of new loss of viable myocardium or new regional wall motion abnormality.

Criteria for prior myocardial infarction

Any one of the following criteria meets the diagnosis for prior MI.

- Pathological Q waves with or without symptoms in the absence of non-ischaemic causes.
- Imaging evidence of a region of loss of viable myocardium that is thinned and fails to contract, in the absence of a non-ischaemic cause.
- Pathological findings of a prior MI.

From Thygson K et al. the Writing Group on behalf of the Joint ESC/ACCF/AHA/WHF Task Force for the Universal Definition of Myocardial Infarction, 2012 Third universal definition of myocardial infarction. *European Heart Journal* 33: 2551–2567, with permission.

CARDIAC MUSCLE STRUCTURE AND BIOCHEMISTRY

Cardiac muscle is found only in the heart; it has much in common with skeletal muscle but is anatomically distinguishable from both skeletal muscle and smooth muscle. Like skeletal muscle, cardiac muscle consists of densely packed bundles of cylindrical muscle cells, multinucleated fibres (~50 µm in diameter and up to several cm in length) which microscopically show characteristic cross-banding or striations. Unlike skeletal muscle cells, cardiac muscle cells are branched and interconnected; they do not have a nerve end plate but are myogenic, initiating contraction without neural control.

All striated muscle (i.e. both skeletal and cardiac, but not smooth muscle) fibres contain many myofibrils, cylindrical bundles composed of two types of contractile protein filaments: thick filaments of myosin (~15 nm in diameter) and thin filaments of actin (~7 nm in diameter). Muscle contraction involves the ATPase-dependent sliding of the thick myosin filaments across thin actin filaments. Each myofibril consists of a chain of contractile units called sarcomeres, each approximately 2.3 µm in length. Electron microscopy reveals that each sarcomere consists of several distinct regions consisting of dark bands alternating with light bands, giving rise to the striated appearance of skeletal and cardiac muscle. The light bands contain only thin actin filaments, whereas the dark bands contain thick myosin filaments overlapping with the actin filaments.

The cyclical interactions between actin and myosin that result in striated muscle contraction are regulated by intracellular calcium concentration, and are initiated by the release of calcium ions from the sarcoplasmic reticulum. The principal proteins which regulate these interactions, tropomyosin and troponin, are located on the actin filament (Fig. 38.3). Tropomyosin is an actin-binding protein which forms a continuous strand lying along the length of the actin filament. Troponin is a tadpole-shaped flexible complex of proteins that lies against the tropomyosin strand. It comprises three subunits:

- troponin T (TnT, 37 kDa), responsible for binding of the complex to the Tropomyosin strand of the thin filament, to form the troponin-tropomyosin (T-TM) complex
- troponin I (TnI, 22.5 kDa), holds T-TM complex in place by Inhibiting the activity of actin-myosin ATPase in the presence of low calcium, thus inhibiting contraction
- troponin C (TnC, 18 kDa), which senses and binds to Calcium and regulates contraction.

The structure of the ternary troponin complex implies that calcium binding to TnC removes the carboxy terminal of TnI from the actin filament. This alters the flexibility and mobility of the troponin complex, forcing tropomyosin away from the actin filaments and exposing the myosin binding site. Myosin cross-bridges then attach on to the actin filament resulting in muscle contraction.

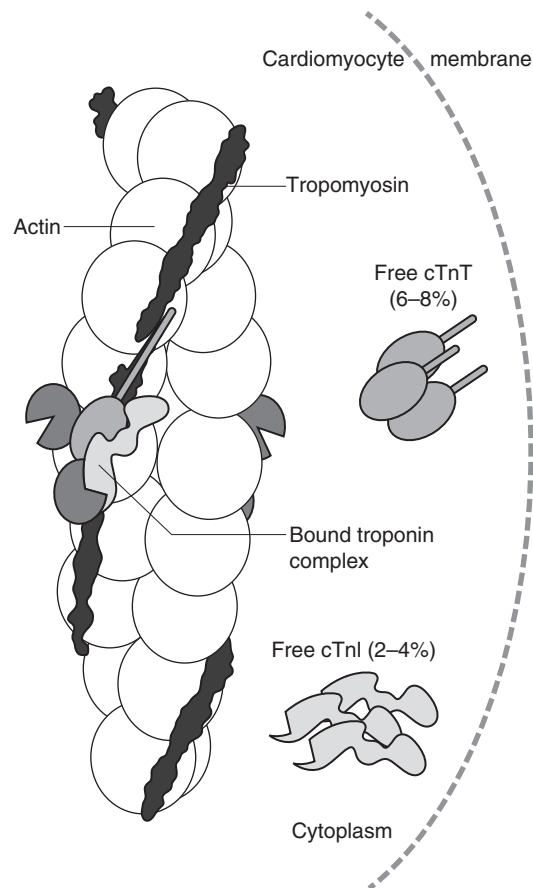


FIGURE 38.3 ■ The thin filament of the cardiomyocyte, consisting of actin with the tropomyosin strand lying along its length and the troponin complex lying against the tropomyosin strand. The three proteins, which constitute the troponin complex (troponin T (TnT), troponin I (TnI) and troponin C (TnC)) are mostly bound to the myofibril, with small amounts of free TnT and TnI in the cytoplasm. From: Gaze D C, Collinson P O 2008 Multiple molecular forms of circulating cardiac troponin: analytical and clinical significance. *Annals of Clinical Biochemistry* 45(4): 349–355, with permission.

Troponin C has only one form, which is distributed throughout all muscles, whereas TnT and TnI each have skeletal muscle and cardiac isoforms, the latter denoted by cTnT and cTnI respectively. Although most intracellular troponin is bound to the thin filament of the myofibril, a small amount of cTnT (6–8%) and cTnI (2–4%) exists free in the cytoplasm.

Effective ventricular pumping depends on effective contraction of the myocardium, the thick muscle that comprises most of the ventricular wall. Immediate energy for muscle contraction is supplied by the hydrolysis of ATP by actin-myosin ATPase, largely generated by the β-oxidation of fatty acids, with oxidative metabolism of ketone bodies and pyruvate generating a lesser amount of ATP. The ATP hydrolysis sites are on the cross-bridges formed between the interacting myosin and actin filaments, and the ATPase is highly active only when they interact in muscle contraction. Depleted ATP concentrations are maintained by the donation of phosphate groups to ADP by phosphocreatine in a rapid transphosphorylation reaction catalyzed by creatine kinase (see Chapter 33). The

myocardium has a rich blood supply and normally extracts approximately 75% of the oxygen from blood passing through the left ventricle. This is facilitated by myoglobin, a cytoplasmic haem-containing protein, which is found in striated muscle and has a high affinity for oxygen. Cardiac muscle cells generate most ATP from the β -oxidation of fatty acids. They are highly dependent on aerobic respiration and do not function well when oxygen supply from coronary arteries is critical.

ARTERIAL STRUCTURE AND FUNCTION

The walls of medium-sized arteries, such as the coronary vessels consist of three cellular compartments separated by two layers of condensed elastic connective tissue (Fig. 38.4). The innermost tunica intima is separated from the media by a thin layer of elastic tissue called the internal elastic lamina. The media is composed of spirally arranged smooth muscle cells which maintain arterial tone and determine the luminal diameter. The tunica media is separated from the tunica adventitia by the external elastic lamina, which is a loose assembly of connective tissue and fibroblasts surrounding the artery. The intima is lined by a single-celled layer of endothelium, which separates the constituents of blood from the arterial wall. The endothelium appears to play a critical role in the regulation of vascular tone, as well as inhibition of leukocyte adhesion and platelet aggregation, through the release of mediators such as nitric oxide (NO) and prostacyclin. Endothelial cells also express tissue plasminogen activator (tPA) and plasminogen activation inhibitor-1 (PAI-1), thus controlling the relative balance between prothrombotic and fibrinolytic activity.

ATHEROSCLEROSIS

Atherosclerosis is a chronic, degenerative, inflammatory condition affecting medium-sized and large arteries. (Atherosclerosis does not normally affect veins, although, when they are used as arterial conduits, e.g. as coronary artery bypass grafts, atherosclerosis may develop very rapidly.) It involves the slowly progressive deposition of lipid and matrix proteins in the arterial wall, which eventually causes narrowing of the lumen. Clinical features

of atherosclerosis are absent until an advanced stage (after several decades) when there is a critical reduction in blood supply to the affected tissues, causing ischaemia or infarction. The acute clinical events associated with atherosclerosis are in most cases due to lesions that are unstable and liable to rupture, leading to haemorrhage into the atherosclerotic plaque. The resultant thrombosis, the formation of an occluding clot within an artery, is the usual cause of myocardial infarction.

Theories of early atherogenesis

Athere is the Greek word for 'gruel', and describes the appearance of the contents of the advanced plaque. Understanding of the cellular events that lead to coronary atherosclerosis has grown substantially over the last 2–3 decades due to: (1) application of cell-specific monoclonal antibodies for immunocytochemical analysis of human lesions; (2) the cloning and characterization of the genes of several cytokines, growth factors and cell surface receptors that may be involved in atherogenesis and that have allowed studies of gene expression, and (3) the use of experimental models including the development of gene knock-out animals.

The response-to-injury hypothesis

Ross first proposed his response-to-injury hypothesis in the early 1970s, focusing on the role of platelets as a possible source of growth factors, and their interaction with the damaged artery wall. With the subsequent realization that the endothelial layer remains intact until the stage of a very advanced ulcerated plaque, the theory was refined to emphasize the importance of more subtle forms of injury to the endothelium, without gross anatomical defects. It was proposed that a number of different agents could contribute to endothelial injury, including smoking, hypertension, hyperlipidaemia and viral infection (Box 38.2). This view has been supported by studies using measures of endothelial dysfunction.

The lipid oxidation hypothesis

The lipid oxidation hypothesis of Steinberg and colleagues provided a further mechanism of endothelial injury, as well as an explanation for the formation of macrophage-derived foam cells that are characteristic of the early lesions of atherosclerosis (Fig. 38.5).

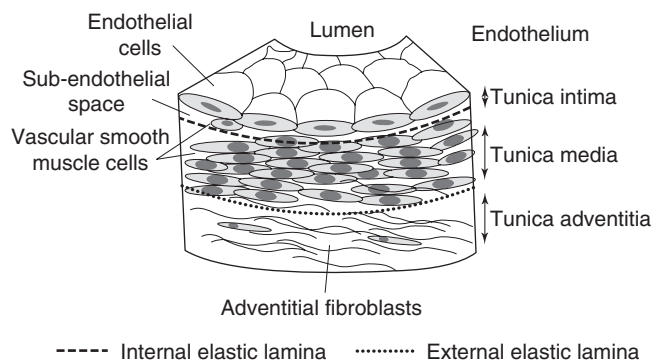


FIGURE 38.4 ■ Structure of a normal medium-sized artery.

BOX 38.2

Potential causes of endothelial dysfunction

- Elevated/modified LDL
- Cigarette smoking
- Hypertension
- Diabetes mellitus
- Genetic
- Infectious microorganisms
- Elevated plasma homocysteine

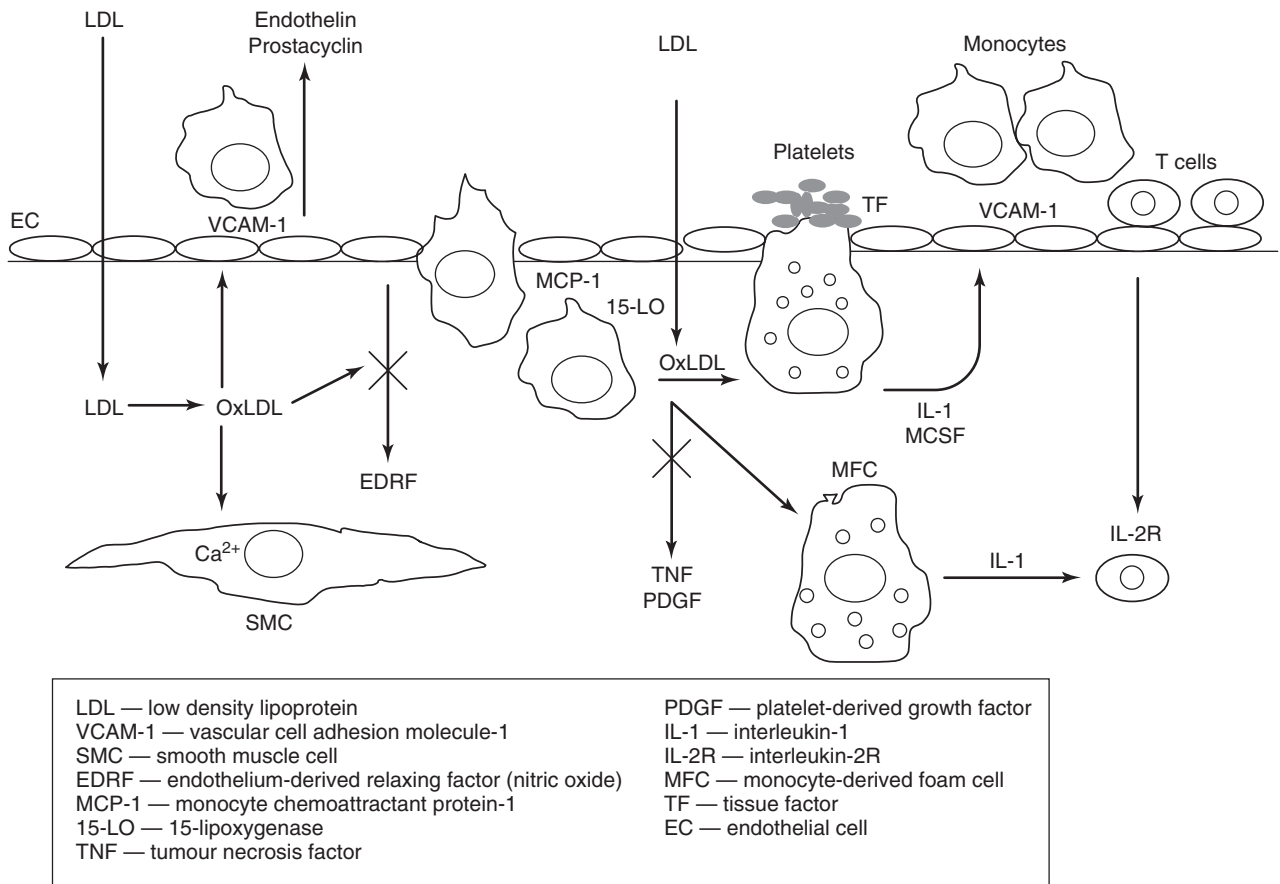


FIGURE 38.5 ■ Steps in the lipid oxidation hypothesis of atherogenesis proposed by Steinberg and colleagues. Adapted from Steinberg, D, Parthasarathy S, Carew T E, Khoo J C, Witztum J L 1989 Beyond cholesterol. Modification of low density lipoprotein that increases its atherogenicity. *N Engl J Med* 1989; 320: 915–924.

Monocytes adhere to the endothelium and accumulate within the subendothelial space at an early stage of the process. Together with smooth muscle cells in the arterial wall, they take up cholesterol in an unregulated manner and are converted to lipid-laden foam cells within the arterial wall, giving rise to the first macroscopically evident lesion, the fatty streak. However, such excessive cellular uptake of cholesterol cannot occur via the low density lipoprotein (LDL) receptor pathway because of the tight regulation of LDL receptor expression by intracellular cholesterol. Goldstein and Brown proposed the existence of the 'scavenger receptor', which allows the unregulated uptake of cholesterol in the form of modified LDL. Several scavenger receptors have now been identified by molecular cloning, the most important of which appears to be CD36 (also called scavenger receptor B). Uptake by these receptors requires chemical modification of the LDL particle by enzymatic, non-oxidative alteration; oxidation, which accelerates the accumulation of cholesterol; glycosylation, or glycoxidation. The oxidation process modifies a lysine amino acid on the apolipoprotein B. Oxidation of LDL can occur in any of the cells within the artery, including the endothelial cells, macrophages, smooth muscle cells and T lymphocytes.

The oxidation of LDL results in the formation of isoprostanes which are chemically stable, free radical-catalysed products of arachidonic acid that are structural isomers of conventional prostaglandins. They reflect

lipid peroxidation and are markers of oxidant stress in hypercholesterolemia and atherosclerosis. Levels of isoprostanes are increased in atherosclerotic lesions and localize to foam cells and the extracellular matrix.

Oxidized LDL particles promote atherosclerosis via one or more of the following effects:

- chemoattractant for monocytes
- promotion of inflammatory and immune changes via cytokine release from macrophages and antibody production
- unregulated uptake via the scavenger pathway leading to foam cell formation (foam cells can rupture, releasing oxidized LDL, intracellular enzymes, and oxygen free radicals that can further damage the vessel wall)
- induction of apoptosis of vascular smooth muscle and human endothelial cells, which suggests a mechanism for the response to injury hypothesis of atherosclerosis
- disruption of the endothelial cell surface which impairs endothelial function, reducing the release of nitric oxide (NO), which is a major mediator of endothelium-dependent vasodilation; damage to the endothelium also promotes platelet adherence and the release of cytokines that stimulate smooth muscle proliferation
- causes an increase in platelet aggregation and thromboxane release, which contributes to vasoconstriction and intravascular thrombus.

There is persuasive evidence supporting the notion that LDL oxidation occurs *in vivo*. Lines of argument include: (a) epitopes of oxidized LDL have been demonstrated in atherosclerotic lesions; (b) LDL isolated from lesions has similar properties to oxidized LDL, cross-reacts with antisera raised against malondialdehyde- and 4-hydroxynon-enal-modified LDL, and is recognized by the scavenger receptor; (c) LDL modification also appears to lead to the expression of neo-antigens that elicit an autoimmune response – autoantibodies to oxidized LDL have been found in human plasma, and within human atherosclerotic lesions.

Inhibiting LDL oxidation *in vivo* by treatment with antioxidants such as vitamin E inhibits experimentally induced atherogenesis; vitamin E can also reduce the uptake of oxidized LDL *in vivo* by reducing expression of the CD36 receptor. Severity of atherosclerosis in several different animal models (rabbit, monkey, hamster, mouse) can be significantly ameliorated by treatment with a variety of antioxidant compounds. Large clinical trials of antioxidants (vitamin E or β -carotene) in the general population have, however, been singularly disappointing, with no effect on cardiovascular outcomes demonstrated by meta-analyses. Rather than the lipid oxidation theory not being relevant to human atherogenesis, this finding may be due to the wrong antioxidant being chosen, the wrong dose, the wrong subjects or the wrong stage of atherogenesis being targeted: foam cell formation is the earliest stage of the atheromatous plaque, and cardiovascular outcomes are noted with advanced plaque formation, generally several decades later.

The fibrofatty lesion

The mature plaque is characterized by a fibrous cap, composed of smooth muscle cells and extracellular matrix, overlying a pool of lipid, cholesterol crystals and inflammatory cells. The conversion of the fatty streak (Fig. 38.6A) into a fibrofatty plaque (Fig. 38.6B) necessitates the recruitment and proliferation of vascular smooth muscle cells. This process is driven by the synergistic interplay of several growth factors, such as platelet-derived growth factor, insulin-like growth factor and basic fibroblast growth factor. These factors are smooth muscle cell mitogens and/or chemotactic factors and are likely to be important contributors to the process.

The complicated plaque/plaque rupture

Unstable plaques, prone to fissuring and rupture, are characterized by a large lipid pool, thin fibrous cap, fewer smooth muscle cells and larger numbers of inflammatory cells (Fig. 38.6C). Activated macrophages within the plaque are a rich source of matrix metalloproteinases (MMPs). These are a family of proteases which, together with other proteases such as cathepsins and elastases, play a key role in tissue remodelling. Under normal physiological circumstances, there is a balance between MMPs, which degrade cardiac extracellular matrix, and tissue inhibitors of metalloproteinases (TIMPs). However, increased MMP activation has been implicated in a wide variety of cardiovascular pathologies, including atherosclerosis. This has the potential to cause localized regions of de-endothelialization that

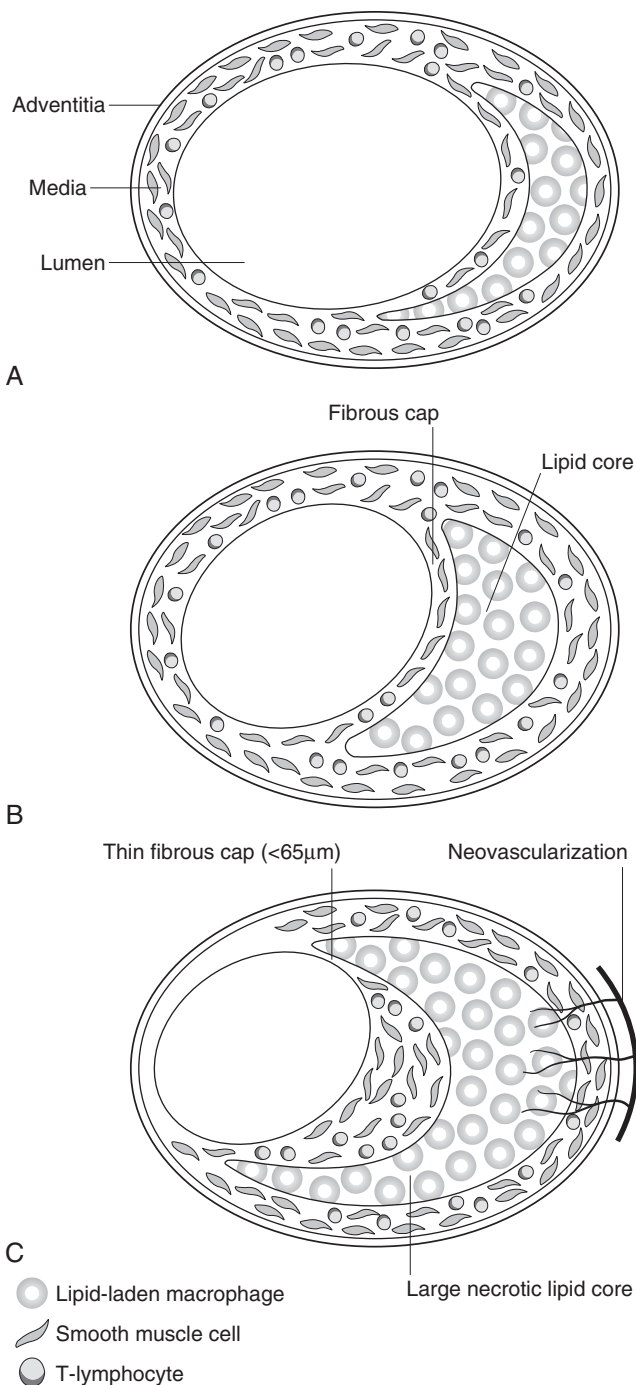


FIGURE 38.6 ■ Features of an atherosclerotic plaque. (A) Fatty streak. Subendothelial collection of lipid-laden macrophages. Plaque grows in an outward direction so that the lumen is conserved. (B) Mature plaque/fibrofatty plaque, characterized by a fibrous cap consisting of smooth muscle cells and extracellular matrix, e.g. collagen. (C) Unstable plaque, characterized by more inflammatory cells and fewer muscle cells, particularly in the shoulder region. Large necrotic lipid core, thin fibrous cap and proliferation of adventitial vasa vasorum.

may destabilize the plaque and lead to focal thrombosis and plaque rupture. The shoulder region of the plaque appears to be particularly vulnerable. In most cases of fatal myocardial infarction, at least one major coronary artery is narrowed by more than 75%, and this is usually associated with plaque fissuring and thrombosis.

ACUTE MYOCARDIAL DAMAGE

Chest pain is responsible for a large proportion of attendances at Accident and Emergency Departments; the spectrum of differential diagnoses is broad and, in addition to ACS, includes muscular pain, which is of little consequence, as well as other potentially fatal diagnoses such as pulmonary embolism. These diagnoses can often be distinguished clinically at the bedside on the basis of the history, examination and ECG findings.

Myocardial ischaemia due to obstructive but stable coronary artery atherosclerosis classically causes pain, discomfort or a feeling of tightness or heaviness in the central chest, left arm or jaw, usually precipitated by exertion and relieved by rest or anti-anginal drugs (e.g. sublingual glyceryl trinitrate). Stable angina results from ischaemia which is of insufficient severity to cause myocardial necrosis; transient ischaemic changes may be apparent on the ECG. Acute coronary syndrome results from erosion or rupture of atherosclerotic plaques in the coronary arteries, with release of the prothrombotic contents of the plaque and superimposed platelet aggregation and thrombosis. The pain is more severe, lasts longer than that of stable angina and does not improve with rest. Ischaemic changes on the ECG may be seen. Complete occlusion of a coronary vessel results in MI with classical ST elevation, i.e. STEMI.

However, history and ECG are not always helpful in differentiating causes of chest pain. A significant proportion of ACS patients do not describe classical pain, and many patients with non-cardiac pain describe features which may suggest cardiac pain. Typically ischaemic ECG abnormalities may not be present in ACS. Most patients with ACS have a partial or transient coronary artery occlusion causing myocardial ischaemia without persistent ST elevation, i.e. unstable angina (UA) or NSTEMI. The term 'non-ST elevation ACS' (NSTEMACS) encompasses both UA and NSTEMI; it reflects the fact that both conditions are considered to be on the same spectrum of myocardial injury, sharing a common pathogenesis but differing in severity and prognosis. The principal feature distinguishing NSTEMI from UA is biochemical evidence of myocyte necrosis.

The importance of accurately and rapidly distinguishing ACS from non-cardiac chest pain is that early treatment of ACS significantly influences mortality and morbidity. Early exclusion of ACS will also help reduce costs of unnecessary hospital admission and investigations, as well as reducing the patient burden.

The relative frequency of the diagnoses in patients admitted with an ACS is approximately: STEMI 30–33%, NSTEMI 25% and UA 38–42%. In STEMI patients, immediate thrombolysis (either with drugs or PCI) is indicated as soon as possible after diagnostic ECG is performed, without waiting for confirmatory elevated serum biomarkers. Patients with NSTEMACS benefit from potent antithrombotic treatment (e.g. low molecular weight heparin/platelet inhibition). The TACTICS-TIMI 18 trial (Fig. 38.7) has indicated that the NSTEMI subgroup may benefit from early coronary angiography and revascularization, whereas those with UA do not, although other studies have not confirmed this benefit and more work needs to be done.

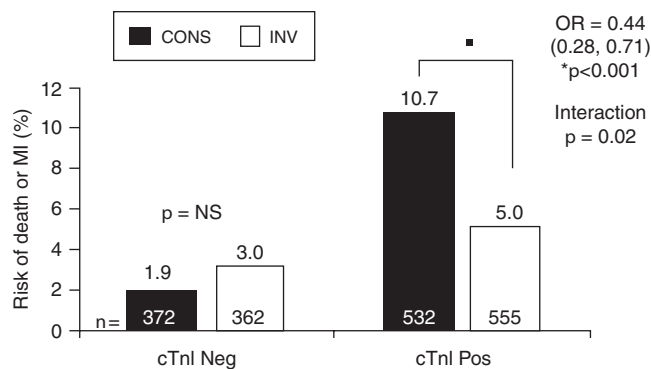


FIGURE 38.7 ■ TACTICS-TIMI 18 trial, indicating improved outcome when patients with NSTEMI and elevated troponin (cTnI Pos) were treated with early invasive intervention (INV) compared to conservative management. The same benefit was not observed in NSTEMI patients without a raised troponin (cTnI Neg). Data from Morrow D A et al. 2007 National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines, Clinical Characteristics and Utilization of Biochemical Markers in Acute Coronary Syndromes. Clinical Chemistry 53: 552–574, with permission.

Nonetheless, it is clear that rapid and accurate diagnosis of ACS is key to improved outcomes. In both STEMI and NSTEMI, damaged myocytes release structural proteins into the circulation, which can be measured and used as biomarkers of acute myocardial injury. The precise definition of myocardial infarction, developed by the 2007 ECC/ACC/AHA/WHF Global Task Force, bases its first criterion on elevations of cardiac biomarkers (see Box 38.1). The current definition gives a preference for use of cardiac troponin, although there is an acknowledgement that the definition is likely to evolve with further scientific advances.

Biomarkers of acute myocardial damage

Within seconds of the onset of myocardial ischaemia, aerobic metabolism ceases within the myocyte, anaerobic glycolysis is initiated and potassium begins to leak out of the cell. This is followed within minutes by leakage of metabolites, a fall in pH and increase in intracellular calcium. Within hours, with sustained depletion of ATP, myocyte necrosis results in irreversible ultrastructural changes. These changes include disruption of the sarcolemma and leakage of macromolecules such as cTnT and TnI, CK-MB and myoglobin.

Since the development of thrombolytic therapies for acute myocardial injury, which must be administered early to be effective, it has become critical to be able to detect myocyte necrosis promptly and accurately. Throughout the 1980s and 1990s, the gold standard for diagnosis of MI was the cardiac-specific isoform of creatine kinase (CK-MB). However, it became apparent that serum CK-MB was not elevated in all cases of myocardial injury: biopsy of myocardium taken during coronary artery bypass surgery for unstable angina showed areas of myocardial micronecrosis, which had not been accompanied by a CK-MB rise. Thus, there has been a drive to develop more specific and sensitive biomarkers of cardiac necrosis, and there is a range of markers currently available, with varying tissue specificity, sensitivity and pattern of release into the circulation (Fig. 38.8).

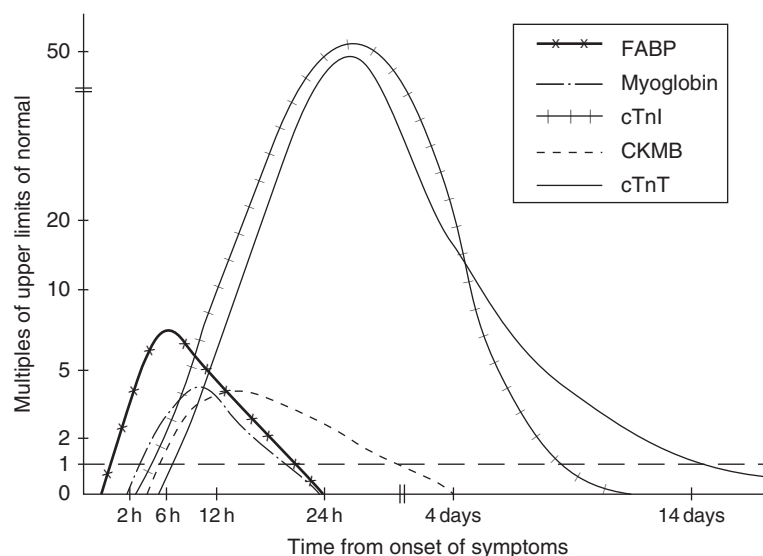


FIGURE 38.8 ■ Pattern of release of some biomarkers into the circulation after acute myocardial infarction.

The pattern of release of biomarkers is influenced by:

- subcellular location: soluble cytosolic molecules are released more rapidly than molecules bound to the myofibrils
- molecular mass: smaller molecules can enter the circulation directly via the microvascular endothelium but are also cleared through the glomerulus
- cytosolic enzymes: increased intracellular calcium activates cytosolic enzymes, including caspase, which promotes dissociation of structurally bound proteins such as cTnT and cTnI.

The ideal biomarker of myocardial injury has a detectable increase in the circulation early, is specific for myocyte necrosis and is not elevated in non-cardiac conditions. None of the currently available cardiac biomarkers meet these criteria, but awareness of their limitations considerably improves clinical utility. The current marker of choice is cTn, with little to choose between cTnI and cTnT.

Troponins

Over the past 20 years, cardiac troponin (cTn) measurement has revolutionized diagnosis of patients with suspected acute coronary syndromes. Monoclonal antibody-based assays specific for the cardiac isoforms of TnT and TnI have been developed that reliably identify myocardial micronecrosis even when CK-MB is not elevated. Recognition of their superior clinical value led in 2000 to the new universal definition of MI based on the elevation of cardiac troponins, updated in 2007 and 2012 to include evidence-based clinical and analytical guidance on use of biomarkers in management of ACS. Cardiac troponins now play a pivotal role in the diagnosis of AMI as well as an important role in the risk stratification of ACS. The current definition of AMI requires detection of a rise and/or fall of cTn (ideally), with at least one measurement >99th centile of a reference population, together with clinical or ECG evidence of ischaemia.

Although most usually found in the myofibril, a small proportion of total cTn also exists in a cytosolic pool (approximately 6% of cTnT and 3% of cTnI) (see Fig. 38.3). The biphasic release of cTnT from damaged heart muscle probably reflects this distribution: the initial release into the circulation following membrane damage during severe ischaemia becomes detectable at 3 h with a peak at 14 h. This is followed by slow dissociation from and degradation of myofilaments, leading to continuous release over 3–5 days. Elevations persist for up to ten days, permitting late diagnosis of MI. There is some debate over whether cTn is released before cell death occurs, i.e. at a reversible stage of injury. In support of this is the finding of transient minor elevations in circulating cTn in some triathletes following extreme exercise. However, in practical terms this is not a critical issue, since, in the majority of situations apart from extreme exercise, increased cTn is associated with a poorer outcome regardless of whether the injury is perceived to be reversible. As well as being the basis for diagnosis of MI, cTn is also helpful for risk assessment and management of suspected ACS patients. Patients previously classified as UA on the basis of normal CK-MB, may be reclassified as NSTEMI, with increased serum concentrations of troponins associated with higher risk of recurrent cardiac events.

In some cases, cardiac TnT antibodies also cross-react 0.5–2.0% with skeletal muscle TnT, thus skeletal muscle may be a source of raised plasma levels. TnI is thought not to be expressed at all in skeletal muscle and may therefore be more specific for myocyte necrosis. However, in the majority of clinical situations, the specificity of cTnT is comparable to cTnI.

There is only one commercially available cTnT assay, but a range of immunoassays for cTnI, which exhibit different analytical sensitivities and specificities. Potential causative factors include a lack of standardization, the occurrence of post-translational modifications of both cTnI and cTnT, and variation in cross-reactivities of the antibody to the various detectable forms of cTnI resulting from its degradation.

The ESC/ACC consensus document on the definition of myocardial infarction recommends that individual laboratories use a cut-off value equal to the 99th percentile of the reference population. At present, cTnT and cTnI are normally undetectable in healthy subjects using conventional assays; therefore, the 99th centile is very low and most assays do not have good precision at this low concentration. This limits the capacity to detect change with serial samples but does not significantly increase the number of false positive results. Thus, in an appropriate clinical scenario, any elevation of cardiac troponin T or I concentration above the 99th centile for a reference population indicates myocardial infarction. The absolute cut-off varies with the assay used, although for cTnT (for which assay there is currently only one manufacturer), this value is often taken to be 0.1 µg/L. It is important that clinicians are familiar with the 99th centile value of the assay used in their own laboratory and are aware that decision limits will vary with different assays. A regularly updated table showing the characteristics of commercially available cTn assays is maintained by the International Federation of Clinical Chemistry and Laboratory Medicine and can be found at: <http://www.ifcc.org/ifcc-scientific-division/documents-of-the-sd/>.

As data have accumulated, it has become evident that even small elevations of troponin are associated with poorer outcomes. In patients with ACS whose troponin concentration is less than the 99th centile, there is still an association between troponin value and clinical outcome. It is thus also useful to define the lower limit of detection. This may be taken to be the lowest concentration at which a between-batch CV of 10% is achievable, and for cTnT this is around 0.03 µg/L. However, most cTn assays in current routine clinical use have poor precision below the 99th centile. The recognition that the assay has clinical utility in this range has led to a continued drive to develop tests which reliably detect lower concentrations, resulting in several generations of troponin assays.

Although terminology for naming generations of cTn assays has been inconsistent, contemporary assays are generally referred to as 'sensitive' assays. The next generation of assays should be able to measure cardiac troponin with a total CV of <10% at concentrations significantly lower than the 99th centile of the normal reference population. Thus they are likely to be termed 'high sensitivity' cTn (hs-cTn) assays. They should reliably measure troponin in most normal individuals and be able to detect changes in serial measurements in the range below the 99th centile, potentially resulting in earlier diagnosis and better management of ACS.

High-sensitivity troponins

The newest generation of Tn assays have a limit of detection that is of the order of 10–100-fold lower than the 'sensitive' assays currently available. By definition, these hs-cTn assays measure cTn in the majority of normal healthy subjects with acceptable precision. They have the potential to increase the promptness and accuracy of diagnosis of ACS. There are a number of different hs-cTn assays in various stages of development; in the case of hs-cTn, a value >14 ng/L is considered abnormal. Separate

decision limits are required for different hs-cTnI assays in the absence of standardization.

Because small differences can be very significant when dealing with such low values, hs-cTn assays are more subject to analytical factors than sensitive assays. Even mild haemolysis reduces hs-cTn values, and sample type (heparin vs EDTA vs serum) can also influence the result. There is greater potential for increase originating from skeletal muscle with hs-cTn compared with the sensitive assay. Some assays show gender-based differences in the 99th centile value.

Since hs-cTn assays can detect cTn in individuals without apparent myocardial damage, it is possible to establish a reference interval for healthy subjects. Values above the 99th centile (a considerably reduced diagnostic threshold compared to 99th centile of the sensitive assays) are associated with a poorer prognosis, and emphasis on detection of a rise and/or fall is likely to be important. Early studies indicate that hs-cTn is likely to facilitate the early diagnosis of AMI as it is presently defined. However, the trade-off for increased diagnostic sensitivity is a reduction in diagnostic specificity. The proportion of patients falsely diagnosed with an MI increases if the diagnosis is based on hs-cTn, particularly if clinical symptoms or ECG changes are uncertain. Even within the reference interval, it is apparent that the higher the value, the greater the risk to the individual, and hs-cTn thus appears to be a better marker for stratifying long-term risk in ACS compared with sensitive assays.

There remain many unanswered questions relating to the use of hs-cTn assays in ACS, including what degree of change can be called an increase, exactly what an increase means, what are its prognostic implications and whether it can be used to influence management to improve outcomes. Until these questions can be answered, the role of hs-cTn in the acute setting will continue to evolve, and must be interpreted in conjunction with the clinical picture. There may be particular problems if the assay is used unselectively in patients with low likelihood of ACS.

Utilization of hs-cTn measurements may be more useful for risk stratifying patients in settings other than acute coronary syndromes including renal insufficiency, heart failure, cardiac amyloid and the elderly.

Other causes of elevated cTn

cTnT and TnI are sensitive and specific markers of myocardial injury when used as recommended by the current universal definition of myocardial infarction, with a cut-off point of the 99th centile. However, lowering the cut-off point may lead to interpretation of an elevated cTn as due to ischaemia, when the elevation is due to myocardial damage due to other causes. To avoid misdiagnosis, it is important to recognize the wide spectrum of disease states other than ACS that can cause raised cTn. These include disorders that would be in the differential diagnosis of MI such as pulmonary embolism, heart failure and myocarditis. The differential diagnosis is given in [Box 38.3](#). Although increased cTn that is not due to ACS may be perceived to be a confounder of clinical interpretation, careful consideration of the source of

BOX 38.3**Differential diagnosis of increased plasma troponin which is not due to acute coronary syndrome or heart failure****Acute disease**

- Cardiac or vascular
 - Endocarditis
 - Myocarditis
 - Pericarditis
 - Coronary artery spasm
 - Acute aortic dissection
 - Cerebrovascular accident
 - Apical ballooning syndrome
 - Gastrointestinal bleeding
 - Medical ICU stay
- Respiratory
 - Pulmonary embolism
 - Adult respiratory distress syndrome (ARDS)
- Infection
 - Sepsis
 - Viral infection
- Miscellaneous
 - Rhabdomyolysis
 - Kawasaki disease
 - Thrombotic thrombocytopenic purpura
 - Neonates: pre-term or very low birth weight
 - Acute complications of some inherited disorders including Duchenne muscular dystrophy and neurofibromatosis
 - Acute environmental exposure to carbon monoxide or hydrogen sulphide
 - Cocaine

Chronic disease

- Cardiac infiltrative disorders
 - Amyloidosis
 - Sarcoidosis

- Haemochromatosis
- Scleroderma
- Hypertension
- Endocrine
 - Diabetes
 - Hypothyroidism
- Renal impairment

Iatrogenic disease

- Invasive procedure
 - Coronary artery bypass grafting
 - Percutaneous coronary intervention
 - Radiofrequency catheter ablation
 - Repair of congenital defect
 - Heart transplant
 - Lung resection
 - Endoscopic retrograde cholangiopancreatography (ERCP)
- Non-invasive procedure
 - Cardioversion
 - Lithotripsy
- Drugs, esp. chemotherapy
 - Adriamycin
 - 5-Fluorouracil

Myocardial injury

- Blunt trauma
- Extreme endurance exercise
- Venom
 - Snake
 - Jellyfish
 - Spider
 - Centipede
 - Scorpion

the elevation will help the clinician determine its clinical significance. There have been occasional descriptions of falsely elevated cTn as a result of analytical interference, but this is a rare occurrence and most frequently elevated cTn in the absence of ischaemia reflects myocardial damage from another cause. Abnormal results may also occur as a result of microparticle interference in anticoagulated samples.

Creatine kinase-MB (CK-MB)

Creatine kinase (CK) is present in large amounts in both skeletal muscle as well as heart muscle, and is also found in brain. Its lack of specificity limits its effectiveness as a cardiac marker, false positives in trauma or post-surgery patients being a particular problem. The enzyme is formed from two dimers, M and B (each with a molecular weight of 40 kDa), and thus three different isoenzymes are possible: CK-MM, CK-BB and CK-MB. The MB isoenzyme is predominantly found in the heart, comprising about 40% of the CK activity in cardiac muscle, and 2% or less of the activity in most skeletal muscle groups and other tissues.

CK-MB isoenzyme is quickly and easily measured and is widely available; it was the gold standard cardiac biomarker until the discovery of the more specific and sensitive cardiac troponin. Measurement of MB isoenzyme can be undertaken by either measurement of the catalytic activity of the MB fraction after antibody inactivation of MM ('activity' measurement) or by direct measurement of CK-MB by immunological detection using antibodies ('mass' measurement). Most current assays measure CK-MB mass, which is more sensitive than activity. Mass assays also largely avoid detection of macro-CKs (CK linked to IgG and dimers of mitochondrial CK), which can confound diagnosis using assays of CK activity. The presence of macro-CK should be considered when CK-MB is a very high percentage (>20%) of total CK.

Even though CK-MB is largely confined to the heart, there are measurable amounts in skeletal muscle, accounting for up to 20% in some muscle groups, although the overall figure is about 1%. Thus, coexisting skeletal muscle injury (including exercise and rhabdomyolysis) can reduce the specificity of CK-MB. Following MI, plasma CK-MB achieves a peak slightly before total CK and also returns to normal more quickly. CK-MB

rises at about 3–4 h after myocardial necrosis, peaks in 10–24 h, and returns to normal within 72 h. Although potentially helpful in the diagnosis of MI at an early stage, it is less useful for the confirmation of diagnosis of late-presenting MI, since the concentration of enzyme generally returns to within the reference limit by three days. The same property makes CK-MB useful for detection of reinfarction, as cTn does not normalize so quickly; however, further increase of cTn from a high baseline also indicates re-infarction.

CK-MB is outperformed in specificity by contemporary cTn assays, and the newer more sensitive cTn assays are likely to take over any remaining clinical utility for CK-MB.

Myoglobin

Myoglobin is a relatively small (17.8 kDa) haem-containing protein found in abundance in the cytoplasm of striated muscle cells, where its main function is oxygen transport. Myoglobin accounts for about 2% of total muscle protein. Its high tissue:plasma ratio, in combination with its small size, means that it rapidly increases in the circulation after myocyte necrosis. Indeed, it is the earliest indicator routinely available for the detection of both skeletal and cardiac muscle damage, appearing in the circulation within 1–3 h of myocardial necrosis, peaking at 6–9 h, before returning to normal by 24–36 h. Its small molecular size also results in a relatively short time frame for its detection as it is also cleared rapidly by the kidneys.

Myoglobin has long been used as an indicator of muscle damage and has received much attention as a cardiac biomarker because of the rapidity of its release. However, its generalized distribution in all striated muscle limits its specificity as a marker of myocardial damage and its usefulness has declined with the increasing availability of cardiac-specific biomarkers. Its relatively high negative predictive value implies that it may be most useful in the early ruling out of acute MI.

Myoglobin can be measured by radioimmunoassay in serum. However, biomarkers of ACS require a short analytical turnaround time if measurements are to be clinically useful. More rapid measurement is provided by latex agglutination immunoassay or fluoroimmunoassay. Clinical utility of point-of-care measurement of myoglobin in combination with cTn with or without CK-MB mass in excluding ACS is currently being evaluated.

Heart-type fatty acid binding protein (H-FABP)

These are also relatively small (15 kDa) proteins found in large amounts in organs associated with significant fatty acid metabolism, which includes the liver and intestine as well as the heart. There are nine distinct types. The H-FABP isoform is present at high concentration within cardiac myocytes, although it is also found at lower concentration in other tissues such as skeletal muscle, distant tubules of the kidneys, the brain, lactating mammary glands and placenta. Under aerobic conditions, long-chain free fatty acid (FFA) metabolism represents the major source of ATP generated in the myocardium. Free fatty acids are poorly soluble in the aqueous phase

and circulate in plasma mainly bound to albumin. In the myocyte cytoplasm, long-chain FFAs are reversibly bound to H-FABP, which is thought to facilitate their transport to the mitochondrial outer membrane where the process of beta-oxidation is initiated.

The potential of H-FABP as a biochemical marker of myocardial injury has been recognized since it was demonstrated to be released from injured myocardium in 1988. Several studies reported its clinical usefulness as an early marker of acute MI (using the World Health Organization criteria as was standard clinical practice at that time). In general it has been shown to perform at least as well or better than myoglobin: this is probably due to its higher relative concentration in cardiac muscle, myoglobin being found in similar amounts in cardiac and skeletal muscle. After acute MI a rise is detectable in plasma as early as 1.5 h after symptom onset, a peak concentration is reached after 4–6 h and, due to rapid renal clearance, the concentration returns to baseline within 20 h.

The normal basal concentration of H-FABP detectable in plasma is likely to be due to continuous release from damaged skeletal muscle. Although its small size makes it useful as an early marker, it also means that it is normally cleared rapidly by the kidney. In renal impairment the plasma level of H-FABP is markedly raised, rendering interpretation difficult. With normal renal function, sustained high concentrations might be indicative of reinfarction, which may be missed by markers such as cTn, which returns to baseline more slowly.

Modern H-FABP assays rely on monoclonal antibodies that have no cross-reactivity (unlike the earlier assays). Point-of-care testing has recently become available. As a standalone test, current H-FABP assays have not been yet shown to have the requisite specificity or sensitivity to safely diagnose early MI. Further work needs to be done to examine its potential role, which may be as part of a panel with a marker which has a complementary timeframe.

Other

Copeptin, a peptide of 39 amino acids, is the C-terminal part of pro-arginine vasopressin (AVP) and is released together with AVP during processing of the precursor peptide. It is thus a surrogate marker for vasopressin, which is significantly involved in the regulation of the endogenous stress response and is present at elevated concentrations in MI, heart failure and different states of shock. Copeptin concentrations are elevated 0–4 h after the onset of symptoms of acute MI, and may be of value in combination with troponin.

Other potential biomarkers of acute myocardial injury that have been evaluated but where insufficient evidence exists to support routine use include ischaemia-modified albumin (IMA), glycogen phosphorylase isoenzyme (GP-BB) and soluble CD40 ligand, among others.

Tests for other causes of chest pain

Measurement of D-dimer may be of some use in the diagnosis of pulmonary embolism. D-dimer is a degradation product released into the circulation when fibrin

(involved in thrombus formation) undergoes endogenous fibrinolysis. A low plasma concentration in the setting of a low clinical risk of pulmonary embolism excludes the diagnosis of pulmonary embolism. However, if the clinical risk is high then the patient needs further investigation (e.g. ventilation-perfusion scanning, CT pulmonary angiography) even if the D-dimer concentration is low. Elevated concentrations of D-dimer occur non-specifically in other conditions, for example MI, pneumonia and other forms of sepsis.

HEART FAILURE

Heart failure (HF) is a failure of the heart to fill with (diastolic HF) or to eject (systolic HF) blood, or both. The term 'congestive heart failure' implies congestion in the lungs due to back pressure caused by failure of the left ventricle to pump blood around the body. By far, the commonest cause is ischaemic heart disease, but it may also result from any structural or functional cardiac pathology, including valvular heart disease, hypertension or viral cardiomyopathy. Dyspnoea is the predominant symptom, although chronic condition heart failure may be relatively asymptomatic in the early stages. It may also present as a medical emergency (acute life-threatening pulmonary oedema). There is a broad spectrum of severity between these extremes and symptomatic heart failure is a relatively common cause of attendance to emergency departments.

Heart failure has an overall prevalence of approximately 2%, although this is markedly higher in the elderly, affecting up to 15% of those aged over 85 years. It is likely to increase in prevalence with 'ageing of the population'. It is more common in men than women at all ages. The condition is progressive, and patients with heart failure have a shortened life expectancy and impaired quality of life. In the UK, over 40% of patients do not survive 18 months from the time of diagnosis.

Both mortality and morbidity can be improved with appropriate treatment (e.g. ACE inhibitors and certain β -blockers), which relies on accurate diagnosis. Clinical uncertainty is known to contribute to poorer outcome in patients presenting with dyspnoea. However, diagnostic certainty is not always straightforward. Two-dimensional and Doppler echocardiography are widely used and in addition to demonstrating left or right systolic or diastolic ventricular impairment may also provide information on aetiology. However, in many healthcare systems, access to such tests may be limited and there has therefore been considerable interest in biochemical tests for screening, diagnosis, prognosis and monitoring of treatment.

Natriuretic peptides

There has been much interest in the natriuretic peptides (NPs) that are secreted by the myocardium, as potential markers for HF. B-type natriuretic peptide (BNP) was originally identified in the brain but is now known to be released primarily by the heart, particularly the ventricles. It is synthesized as a prohormone, the C-terminal of which is cleaved on release from the

myocardium to produce BNP (a 32-amino acid biologically active peptide) and an N-terminal fragment of the prohormone called NT-proBNP (76-amino acid and biologically inert). Both can be measured in plasma, and in normal subjects, the plasma concentrations of BNP and NT-proBNP are similar (approximately 10 pmol/L). However in both systolic and diastolic heart failure, plasma concentrations of both rise, plasma NT-proBNP proportionately more than BNP, with NT-proBNP concentrations approximately four-fold higher than BNP concentrations.

Critical values

Measurement of either BNP or NT-proBNP helps to discriminate between HF and other causes of dyspnoea. A normal value of either (BNP <100 ng/L [29 pmol/L] or NT-proBNP <400 ng/L [47 pmol/L]) makes decompensated heart failure highly unlikely ('rule-out' values) and suggests a respiratory or other cause of dyspnoea. High plasma concentrations (BNP >400 ng/L or NT-proBNP >2000 ng/L) strongly support a diagnosis of abnormal ventricular function ('rule-in values'). Intermediate values (BNP 100–400 ng/L [29–116 pmol/L] or NT-proBNP 400–2000 ng/L [47–236 pmol/L]) should prompt a search for a non-cardiac cause of dyspnoea: for example, COPD. Approximately 75% of patients in this grey zone will have HF, which is relatively mild and has a good prognosis.

Non-HF factors influencing NPs

Measurements vary with the assay used and with age, gender, and body mass index: normal values tend to increase with age and to be higher in women than men and, for BNP, lower in obese individuals. Plasma concentrations of both tend to be less elevated in heart failure with preserved ejection fraction (EF) than in heart failure with low EF. A raised concentration may occur in other conditions such as hypoxia (e.g. pulmonary embolism, chronic obstructive pulmonary disease), myocardial ischaemia, tachycardia, atrial fibrillation, left ventricular hypertrophy, right ventricular overload, renal impairment (particularly NT-proBNP), liver cirrhosis, diabetes and sepsis. It is not necessary to adjust rule-out values for age or gender, but it has been suggested that obese (BMI >30 kg/m²) individuals should have their BNP doubled to use the cut-off points above. To date, no correction has been suggested for NT-proBNP in obesity. Plasma concentrations of both, but particularly NT-proBNP, should be interpreted in conjunction with an estimate of renal function.

Clinical utility

When raised concentrations are due to heart failure they are a useful prognostic indicator. Trials with either of these diagnostic biomarkers suggest that their use may reduce both the length of hospital stay and the overall cost of treatment. Both markers provide prognostic information in patients with acute and chronic HF, and plasma BNP also has prognostic value in patients with asymptomatic or minimally symptomatic LV dysfunction.

Guidance from the UK National Institute for Health and Care Excellence (NICE) (2010) suggests measurement of BNP or NT-proBNP as first line investigation in patients clinically suspected of having HF, where there has not been a previous MI (Fig. 38.9). It also suggests that patients with high concentrations (BNP >400 ng/L or NT-proBNP >2000 ng/L) should be referred within two weeks for echocardiography and specialist advice, because of the implications for poor prognosis. A raised BNP or NT-proBNP is also an independent prognostic indicator of mortality in patients at high risk of CHD and in those with established CHD. The value of serial measurements to guide management of acute HF has not yet been conclusively demonstrated, possibly due to the considerable intra-individual biological and analytical variability.

Work is also currently underway to assess diagnostic and prognostic benefits of mid-regional pro-atrial natriuretic peptide (MR-proANP) in HF.

CARDIOVASCULAR RISK FACTORS

Cardiovascular risk assessment

Assessment of cardiovascular risk is important to allow targeting of preventative measures to those who will achieve the greatest benefit. The most widely used cardiovascular risk prediction equation was developed from data collected in the Framingham Heart Study which was initiated in 1948 in the town of Framingham, near Boston, USA. It set out to identify the factors contributing to coronary disease using a cohort design in patients initially without coronary disease. The Framingham Risk Score has proved remarkably useful in diverse populations considering that it was derived from data on a predominantly middle-class, middle-aged, Caucasian population, and therefore underestimates overall lifetime risk and risk in the elderly. It is reasonably accurate at predicting

Diagnosing heart failure

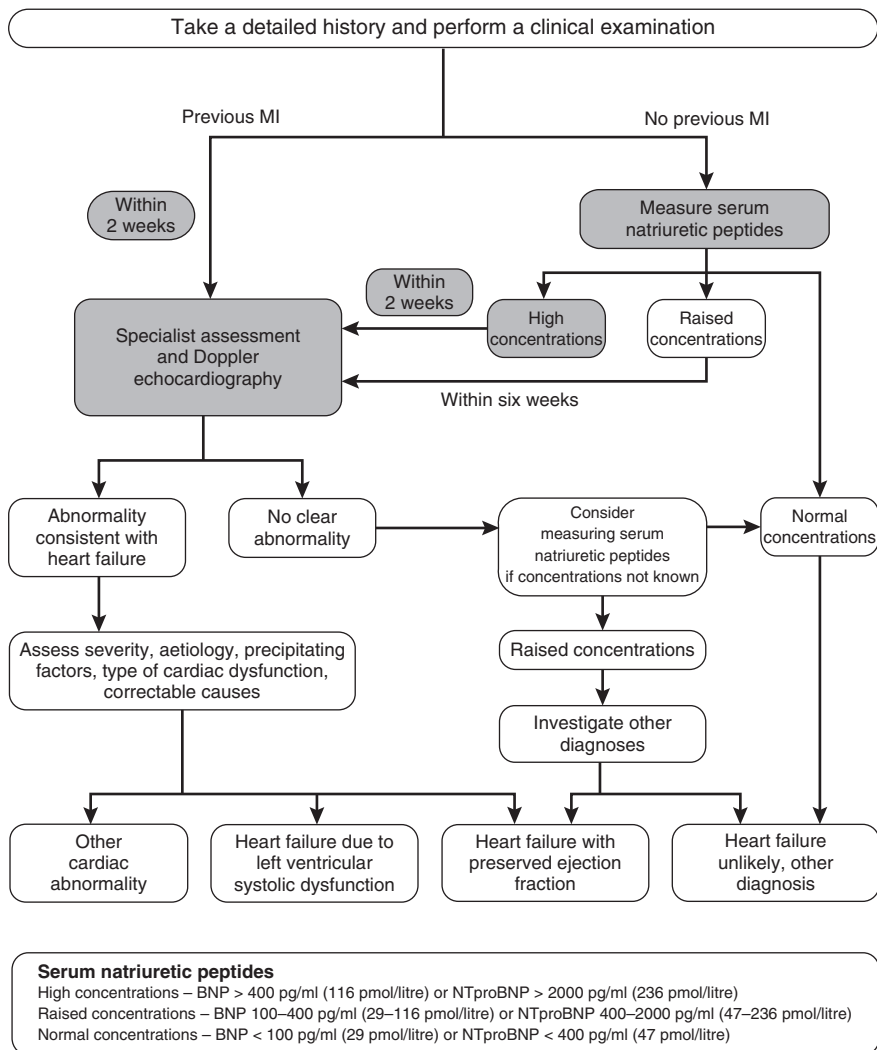


FIGURE 38.9 ■ Diagnosis of heart failure algorithm. From NICE 2010 Clinical Guideline 108 Chronic Heart Failure: management of chronic heart failure in adults in primary and secondary care, with permission.

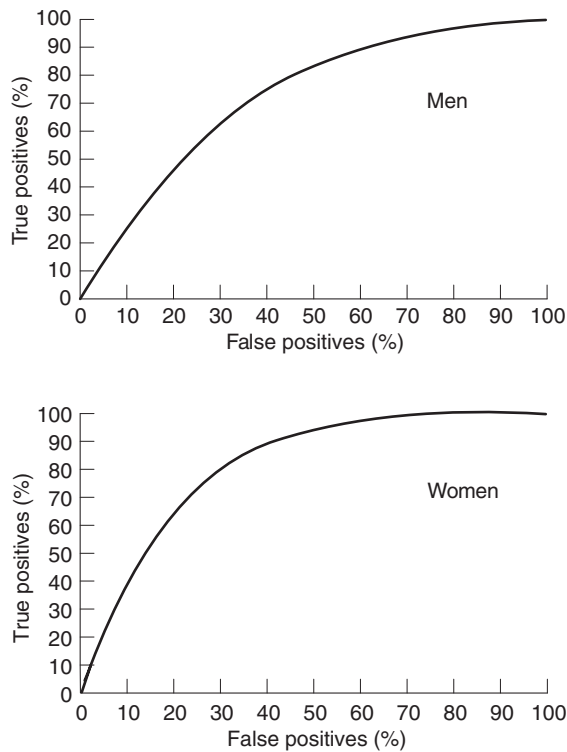


FIGURE 38.10 ■ Receiver operating characteristic (ROC) curves for the Framingham algorithm in men (upper graph) and women (lower graph) using National Health and Nutritional Examination Survey (NHANES) follow-up data. Adapted from Liao Y, McGee D L, Cooper R S, Sutkowski M B 1999 How generalizable are coronary risk prevention models? Comparison of Framingham and two national cohorts. *American Heart Journal* 137: 837–845, with permission.

coronary events, with an area under the receiver operating characteristic (ROC) curve of 0.71–0.76 in males and 0.76–0.81 in females (Fig. 38.10).

The Framingham risk score continues to be used in the UK, although it somewhat overestimates cardiovascular risk in most subjects. In recent years, it has been surpassed in terms of accuracy by the QRISK and QRISK2 scores, which are derived from the UK general practice database and it is likely that QRISK2 or its future iterations will become the generally used cardiovascular risk prediction equation in the UK. QRISK2 employs the same core set of risk factors as Framingham but assigns different weightings to some and incorporates additional risk factors (notably postcode as a marker of social class), which impact significantly on cardiovascular risk in the UK. Alternative risk scores have been developed for use in other populations, notably EUROSCORE, which is widely used across Europe and the Reynolds risk score, which incorporates CRP as an additional risk factor. The major risk factors for cardiovascular disease are similar across all populations, although their relative importance may differ.

Unmodifiable risk factors

Age

Although post-mortem studies have demonstrated that fatty streaks and more advanced lesions are present in the second decade of life, these do not usually

become clinically evident until the fifth or sixth decades. Experimental models of atherosclerosis suggest that some of these early lesions are, to some extent, reversible. Atherosclerosis becomes increasingly prevalent after the age of 20 years. This may be partly associated with the increase in plasma cholesterol concentration and other coronary risk factors with age. The extent of coronary and aortic atherosclerosis increases with age (Fig. 38.11) and risk factor burden (Fig. 38.12). Although the absolute risk of CHD rises with age, the relative risk with increasing cholesterol is steeper in the younger age groups.

Gender

The risk of CHD in premenopausal women is approximately 30% of that in men at any particular age, irrespective of smoking status and the presence of

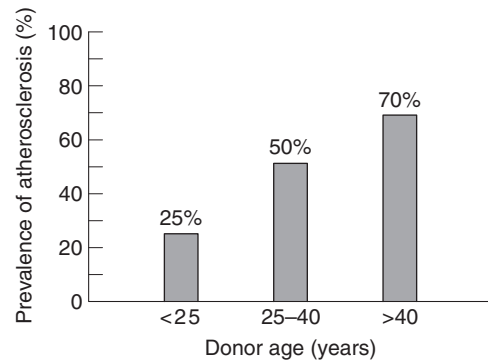


FIGURE 38.11 ■ Prevalence of atherosclerosis identified by intravascular ultrasound in the coronary arteries of heart donors by age. From: Tuzcu E M et al. 2001 High prevalence of coronary atherosclerosis in asymptomatic teenagers and young adults. *Circulation* 103: 2705–2710, with permission.

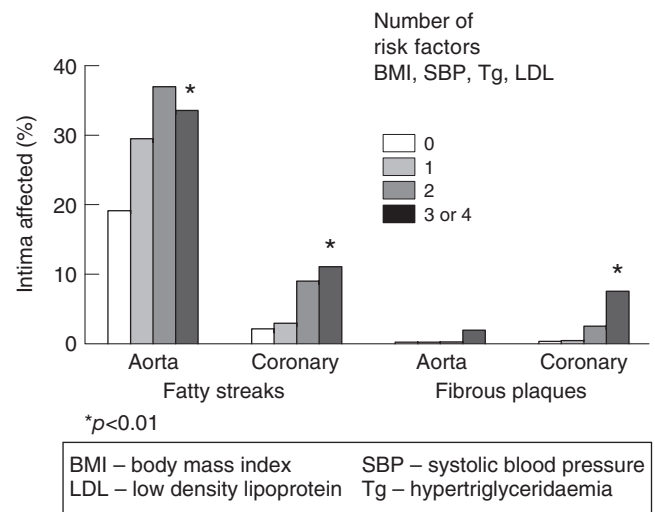


FIGURE 38.12 ■ Effects of cumulative risk factors on coronary and aortic lesion development in children and young adults. Data from Berenson G S et al. 1998 Association between cardiovascular multiple risk factors and atherosclerosis in children and young adults. The Bogalusa Heart Study. *New England Journal of Medicine* 338: 1650–1656, with permission.

hypertension. In women, a greater proportion of cholesterol is present as high density lipoprotein (HDL) cholesterol and this may be protective. Furthermore, it is possible that oestrogens have a direct protective effect on the vasculature. After the menopause, plasma LDL-cholesterol concentrations rise, HDL-cholesterol (HDL-C) falls, there is an increase in visceral adiposity and the risk of CHD rises considerably.

Race

The incidence of CHD and stroke varies with race. In the USA, black people have a higher risk of CHD than white people, while Hispanic groups have lower rates than either group. In the UK, men and women of South Asian descent have a particularly high incidence of CHD (approximately 50% higher rates of premature CHD death compared with the Caucasian population). The difference is widening, as rates do not appear to be falling as fast among this subgroup as for the UK as a whole. This may relate, in part, to a higher prevalence of diabetes in this group. Other factors contributing to this excess mortality may be socioeconomic status, a proatherogenic diet, lack of exercise, enhanced inflammatory status and high plasma concentrations of homocysteine and Lp(a). QRISK2 performs significantly better than the Framingham risk equation for estimating cardiovascular risk in South Asian subjects providing ethnicity is appropriately coded.

Family history

A family history of premature CVD is an important risk factor. In some cases this may be attributable to a single gene effect such as is observed in familial hypercholesterolaemia (see below), although, in most cases, the basis of the underlying genetic predisposition remains unknown and is likely to be polygenic. Family history is a particularly important contributor to risk in men in the lowest quintiles for calculated CHD risk. In women, age-adjusted risk is increased almost three-fold if either parent has CHD before 60 years of age.

Genetic factors

There are a number of inherited traits that are associated with increased CHD susceptibility. Familial hypercholesterolaemia (incidence ~ 1 in 500 in the UK) and familial combined hyperlipidaemia (1 in 100) account for a large proportion of cases. More common genetic variants have a more subtle effect on coronary risk. For example, the apolipoprotein E gene polymorphism is associated with a modest effect on serum cholesterol concentrations, and compared with individuals with the *E-3/E-3* genotype, carriers of the *E-4* allele have a 42% higher risk of CHD, whereas the apo E-2 allele appears not to confer significantly increased risk. The apolipoprotein (a) (apo(a)) gene polymorphism, due to tandem repeat within the kringle 4 region, affects the concentration and molecular weight of plasma Lp(a), and high plasma concentrations of Lp(a) are found more commonly among patients with coronary disease.

Low birth weight

Barker and colleagues have demonstrated that a low birth weight is associated with an increased risk of CVD and type 2 diabetes mellitus. This may reflect a permanent metabolic change resulting from undernutrition during critical stages of early development, which may become manifest as impaired endothelial function later in life. Low birth weight and high body mass in adulthood appear to interact to predict CHD.

Potentially modifiable risk factors

Smoking

The seminal studies of Doll and colleagues were the first to identify the dangers of cigarette smoking with regards to CHD risk and death: approximately 20% of cardiovascular deaths can be attributed to smoking. The injurious agents in cigarette smoke probably include tar, carbon monoxide and free radicals. The assessment of the risk of cigarette smoking is confounded by a number of other variables including social class, ethnicity and gender, but it remains important even when allowance is made for them. The relative risk of CHD in smokers is particularly high in young adults, especially women. Smoking rates differ regionally and, to some degree, contribute to the geographical, racial and socioeconomic differences in CHD rates. Furthermore, smoking also has effects on other risk factors such as plasma lipid concentrations, particularly HDL-C, clotting factors and fasting glucose concentrations. Cross-cultural studies indicate that smoking by itself is insufficient to cause CHD, but increases susceptibility to other risk factors. Passive smoking is also associated with an increased risk of coronary disease (~15–50%). The prevalence of cigarette smoking in the UK has fallen over the past three decades, as has overall tobacco consumption. Smoking cessation is associated with a substantial fall in coronary risk. The benefit of smoking cessation is particularly marked in patients who have already suffered an MI and is associated with a substantial reduction (36%) in mortality in patients with established CHD.

Lipids and lipoproteins

Numerous studies have demonstrated a positive association between CHD and plasma total cholesterol concentration. The association is continuous, exponential and shows no threshold, even at very low concentrations. However, the distributions of plasma total cholesterol concentrations among patients with and without CHD overlap to a considerable degree. Plasma apolipoprotein B (apo B) concentrations appear to be more discriminating: however, they are of less practical utility, as few clinical studies have used apo B concentrations as a basis for treatment (see also Chapter 37).

An inverse relationship between plasma HDL-C concentrations and CHD risk has also been demonstrated in various studies. HDL-C therefore forms an important component of risk prediction equations. However, the results of recent clinical trials using drugs to modify HDL-C levels have generally been disappointing, so

HDL-C is not currently a therapeutic target. Just as apoB may be more discriminating than total or LDL-C, so there is some evidence that apoA1 may be more discriminating than HDL-C.

Over recent years, the importance of triglycerides as a CHD risk factor has received increasing attention. There is a very strong inverse relationship between plasma HDL-C (particularly the HDL₂ fraction) and triglycerides, and, consequently, it has been difficult to demonstrate an independent relationship. However, recent data do support a role for triglycerides as an independent risk factor. Concentrations of plasma triglycerides above approximately 1.7 mmol/L are associated with the formation of the more atherogenic, small, dense LDL. However, triglycerides themselves also have proatherogenic effects by promoting a procoagulant state, being associated with enhanced factor VII activity.

Most primary (genetic) dyslipidaemias predispose to premature CHD, although hyperchylomicronaemia is associated with acute pancreatitis rather than CHD. However, while in the general population approximately 50% of the variability in plasma cholesterol is genetically determined, these monogenic disorders account for only a small proportion of it, and the basis of the greater part remains unknown.

Thrombogenesis, rheology and clotting factors

Thrombosis is the usual terminal event in coronary atherosclerosis. Several studies have revealed the importance of plasma clotting factor concentrations as risk factors for CVD. Fibrinogen forms the substrate for thrombin and represents the final step in the coagulation cascade. It is essential for platelet aggregation, modulates endothelial function and promotes smooth muscle proliferation. A recent systematic review has shown that plasma fibrinogen concentration is moderately strongly associated with CHD, stroke and vascular mortality in middle-aged adults. Fibrinogen concentrations are also related to risk of ischaemic stroke and CHD events in patients who have had a previous stroke. The relationship between plasma fibrinogen and coronary risk may underlie the positive association between plasma viscosity and CHD. However, it is unclear whether these associations are direct, or related to other risk factors such as smoking, so lowering fibrinogen may not itself reduce risk.

Endothelial injury causes the release of tissue factor that, in turn, activates the intrinsic clotting cascade. Platelet activation and aggregation are crucial processes in atherothrombogenesis, and platelet reactivity has been reported to be elevated in subjects with unstable angina and diabetes. The clinical benefit of antiplatelet drugs such as aspirin in patients at coronary risk is now well established.

There is a balance between the formation of clot and its inhibition and dissolution by factors such as proteins C and S and plasmin. The effectiveness of the fibrinolytic system depends on the balance between tissue plasminogen activator (tPA) and inhibitors of plasminogen activation including PAI-1. Tissue plasminogen activator converts plasminogen to plasmin, which acts on fibrin, causing clot dissolution. This process is inhibited by PAI-1, high

concentrations of which are associated with a high risk of reinfarction. Evidence from the Framingham study indicates that plasma PAI-1 concentrations rise with increasing systolic and diastolic blood pressure. Atherosclerotic lesions from diabetic subjects have been shown to contain high concentrations of PAI-1, and plasma PAI-1 is strongly associated with several CHD risk factors including body mass index (BMI), lipids and alcohol intake, and these effects appear to be cumulative.

Apolipoprotein(a) is a glycoprotein that has structural homologies with plasminogen. It is attached to apo B by a disulphide bond and, in some individuals, comprises the major cholesterol rich lipoprotein. High concentrations of Lp(a) have been shown to be associated with an increased risk of CHD, particularly when associated with raised concentrations of LDL-C or homocysteine. The structural similarities between apo(a) and plasminogen have led to the proposition that Lp(a) inhibits plasmin activity, leading to a prothrombotic state. Plasma concentrations of Lp(a) are largely genetically determined, but can be modified, to a limited degree, by dietary fatty acids, oestrogen, the lipid-lowering agent, nicotinic acid and alcohol.

Hypertension

Coronary heart disease risk increases both with increasing systolic and diastolic blood pressure (BP). Blood pressure, like plasma cholesterol, is a continuous variable and there is no clear cut-off value, but hypertension doubles the risk of CHD at any given concentration of cholesterol. Because intra-subject blood pressure measurements can vary so much over time, the diagnosis of hypertension relies on the measurement of blood pressure on several occasions. Blood pressure increases with age and its prevalence varies with ethnicity. Hypertension is also associated with obesity and dyslipidaemia, often as part of the metabolic syndrome.

Obesity

Obesity is an independent risk factor for CHD and is rising in prevalence throughout the developed and developing world (see Chapter 11). Body mass is positively related to fasting triglyceride concentrations, plasma cholesterol and blood pressure, and inversely related to HDL-C. The distribution of body fat appears to be particularly important. It is central or visceral obesity, measured as waist circumference, that is most strongly related to insulin resistance and CHD risk. Waist circumference is a significantly better index of insulin resistance than either waist/hip ratio or BMI. A cut-off value for waist circumference of <100 cm excludes insulin resistance in both sexes with optimal sensitivity and specificity. It has been proposed that hyperinsulinaemia stimulates 11 β -hydroxysteroid dehydrogenase in omental adipose tissue, generating cortisol and promoting a cushingoid fat distribution. Adipose tissue is now recognized to be a source of a number of inflammatory cytokines (interleukin-6, tumour necrosis factor α), growth factors (heparin binding epidermal growth factor) and hormone-like substances (leptin, adiponectin, resistin).

Weight loss is associated with an improvement in a number of coronary risk factors including LDL, HDL, triglycerides, systolic BP and fasting blood glucose. Rapid weight gain in childhood (between 2 and 11 years) appears to predict coronary disease in adulthood.

Impaired glucose tolerance and diabetes

Both diabetes mellitus and impaired glucose tolerance are important risk factors for CVD (Fig. 38.13) and epidemiological data have led to the notion that diabetes confers a similar risk of a cardiovascular event to a prior myocardial infarction (Fig. 38.14). In patients with diabetes, a high haemoglobin A_{1c} (HbA_{1c}) (>10mmol/mol) is associated with an increased risk of CHD; relative risk increases by 1.2-fold for each percentage point increase of HbA_{1c}. However, the value of tight glucose control in preventing macrovascular disease is less clear than its benefits in the prevention of microvascular complications. This may be in part because the atherogenic lipid profile typically associated with type 2 diabetes tends to persist even with excellent glycaemic control.

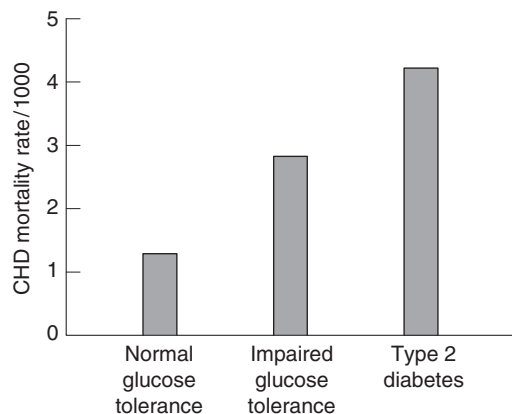
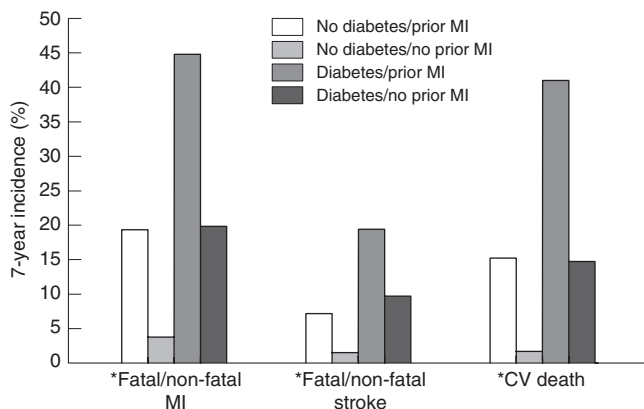


FIGURE 38.13 ■ The effects of glucose tolerance on CHD mortality rates. From: Hsueh W A, Law R E 1998 Cardiovascular risk continuum: implication of insulin resistance and diabetes. American Journal of Medicine 105: 4s–14s, with permission.



* $p < 0.001$ for diabetes vs no diabetes

FIGURE 38.14 ■ Diabetes mellitus confers a similar risk for a cardiovascular event as a prior myocardial infarction (MI). Data from Haffner S M et al. 1998 Mortality from coronary heart disease in subjects with type 2 diabetes mellitus and in non diabetic subjects without prior myocardial infarction. New England Journal of Medicine 339: 229–234, with permission.

Although the absolute CHD mortality is higher for diabetic men than for diabetic women, the gender-related protection in women is lost if they have diabetes.

Metabolic syndrome

The clustering of several coronary risk factors (high triglycerides, low HDL, obesity, hyperuricaemia, hyperinsulinaemia and hypertension) has been known for several decades, and led Reaven and others to propose the existence of a syndrome with a common underlying metabolic defect. The genetic basis for this, if one exists, has yet to be identified. This is further complicated by the many synonyms and definitions in common use. These include definitions developed by the World Health Organization (WHO), International Diabetes Federation (IDF), American Diabetic Association, American Association of Clinical Endocrinologists (AACE), American Heart Association and the National Cholesterol Education Programme Adult Treatment Panel III (NCEP-ATPIII) (Table 38.1). Insulin resistance and hyperinsulinaemia appear to be characteristic features of the syndrome. The latter has also been shown to be associated with coronary events (Fig. 38.15). No matter how defined, there is a high prevalence of metabolic syndrome in westernized populations, with particularly high rates in some ethnic groups. Recent definitions have included ethnic-specific criteria. The prevalence of metabolic syndrome is approximately 22–39%, dependent on the definition. At any particular level of coronary risk, patients with metabolic syndrome appear to have higher event rates than predicted. Furthermore, the metabolic syndrome also predicts the future development of diabetes mellitus, and is associated with a raised concentration of inflammatory markers, such as C-reactive protein (CRP).

Physical activity

A high level of physical activity protects against CHD and probably does so by a number of different mechanisms. For example, it is associated with a reduction in blood pressure and improved lipid profile. Recent data indicate that approximately 30 min/day of moderate exercise is required to have a significant impact on coronary risk. Exercise-based rehabilitation for patients with CHD is effective in reducing total mortality and lipid concentrations. High levels of physical activity also decrease the risk of stroke in a dose-dependent manner. Reduction in fitness predicts cardiovascular death in middle-aged men.

Psychological factors

Data from the mid-1970s indicated that a type A personality, characterized by high achievement and drive, was associated with increased risk of CHD. However, more recent data do not support these findings. Stress adversely affects blood pressure, sleeping patterns and plasma lipid concentrations and there is clearly a plausible interaction between stress and personality. Some studies have demonstrated an increased risk of CHD in patients with depressive illness, though the basis of the association is unclear.

TABLE 38.1 Definitions of the metabolic syndrome/insulin resistance syndrome

	NCEP-ATPIII	WHO	AACE	IDF
Waist circumference (cm)				
Men	>102	–	–	≥94 (≥90 ^a)
Women	>88	–	–	≥80 (≥80 ^a)
Body mass index (kg/m ²)	–	≥30	≥25	–
Triglycerides (mmol/L)	≥1.7	≥1.7	≥1.69	≥1.7
HDL cholesterol (mmol/L)				
Men	<1.0	<0.9	<1.04	<1.04
Women	<1.2	<1.0	<1.29	<1.29
Blood pressure (mmHg)	≥130/85	≥140/90	≥130/85	≥130/85
Fasting blood glucose (mmol/L)	≥6.1	≥6.1	≥6.1	≥5.6
Blood glucose 2-h post challenge	–	–	≥7.8	Diabetes or impaired glucose tolerance
Urinary albumin:creatinine ratio (mg/g)	–	>30	–	–
Other factors			Polycystic ovary syndrome; family history of type 2 diabetes, hypertension or CVD	

^aFor South Asians and Chinese.

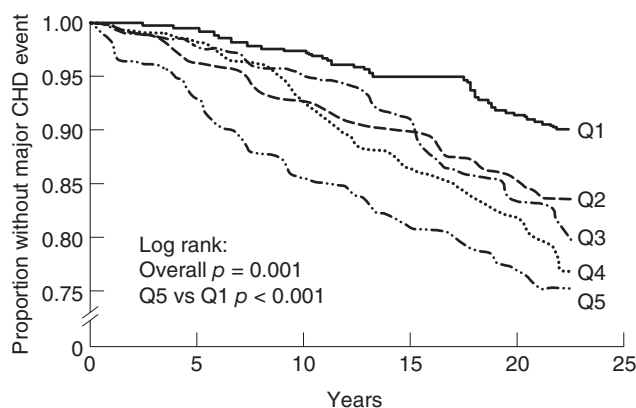


FIGURE 38.15 ■ High plasma insulin concentrations predict increased CHD events in non-diabetic subjects (Q1-Q5 are quintiles of insulin concentration: subjects with the highest concentrations are in Q1, those with the lowest in Q5). From: Pyörälä M et al. 1998. Hyperinsulinaemia predicts coronary heart disease risk in healthy middle-aged men. *Circulation* 98: 398–404, with permission.

Inflammation and infection

Atherosclerosis bears many hallmarks of a chronic inflammatory disease, and at every stage of its evolution is characterized by macrophage and T lymphocyte infiltration. The possible stimuli for this inflammatory process include oxidized LDL, homocysteine, free radicals generated from cigarette smoking and infectious microorganisms. If the original insult were not adequately neutralized, the inflammation might persist, causing the local and systemic release of growth factors and cytokines. These can cause intimal thickening by stimulating smooth muscle cell migration, proliferation and extracellular matrix elaboration. The release of IL-1 β and IL-6 from activated leukocytes may also lead to an induction of hepatic CRP synthesis.

Over the past few years, there has been an increasing interest in the use of inflammatory markers to estimate the risks of acute events in patients with established

coronary disease. In part, the predictive value of these markers may be related to their ability to identify patients with vulnerable plaques, which are rich in activated leukocytes. The risk associated with a high plasma CRP concentration is stronger than that associated with raised von Willebrand factor or erythrocyte sedimentation rate (ESR) but is estimated to be weaker than an increased cholesterol or positive smoking habit. It has also been reported that plasma concentrations of CRP are positively associated with subclinical carotid and femoral atherosclerosis, and a number of established coronary risk factors in healthy, middle-aged men. Among these factors were smoking habit, indices of adiposity, blood pressure, triglycerides and HDL. Of these, smoking habit showed the strongest association with plasma CRP. C-reactive protein is deposited in lipid rich regions of human coronary arteries, and its accumulation may precede monocyte recruitment. It is chemotactic for monocytes, and may also be involved in the activation of complement. A relationship between plasma concentrations of CRP and basal endothelial cell nitric oxide (NO) synthesis has also been reported, suggesting that CRP may indeed be involved early in the pathogenesis of atherosclerosis.

It remains unclear whether CRP is a risk marker or risk factor for atherosclerotic cardiovascular disease. Mendelian randomization studies suggest that genetic variation associated with increased CRP levels is not associated with increased cardiovascular risk, suggesting that CRP is a risk marker. However, it is now recognized that a range of conditions associated with chronic inflammation and elevated CRP, such as rheumatoid arthritis and psoriasis, are associated with a significant increased risk of atherosclerotic vascular disease.

Relative importance of coronary risk factors

Although many risk factors for CHD have been described, and others no doubt will become recognized in the future, the contribution of many of them is relatively

minor, certainly in comparison with diabetes, cigarette smoking, hypertension and hypercholesterolaemia. It is to these risk factors that the greatest effort should be directed in seeking to ameliorate an individual's risk of CHD.

Dietary factors

Diet is known to modulate several established coronary risk factors, and dietary change is an important element of coronary risk management as outlined by various dietary guidelines. These are targeted at optimizing the lipid profile, attaining ideal body weight and reducing blood pressure.

Salt

In experimental models, an increased dietary intake of sodium chloride raises blood pressure. This appears to be confirmed by clinical studies such as the INTERSALT study. Sodium excretion has been found to be related to median systolic and diastolic BP. A 5.9 g difference in intake of salt was associated with a 3–6 mmHg difference in blood pressure. A reduction of dietary salt causes a significant reduction in blood pressure, although sensitivity to salt intake varies between individuals. Up to 50% of hypertensive subjects are sensitive to salt loading, an effect that varies with ethnicity. The intake of other minerals also appears to modify blood pressure; these include potassium, calcium and magnesium, the intakes of which are inversely related to blood pressure.

Simple sugars

Dietary simple sugars (sucrose, lactose, glucose and fructose) comprise approximately 25% of total energy intake. An association between sucrose intake and CHD risk was first proposed in the 1960s, but this has not been supported by recent studies. Carbohydrates differ in their effects on blood sugar and insulin concentrations. Low glycaemic index (GI) foods have less effect on postprandial blood glucose and insulin. The Nurses' Health study has shown that a high glycaemic load is associated with an increased risk of CHD. This may be related to adverse effects on HDL and triglycerides, and increased risk of diabetes. However, low-GI diets appear to have only a modest impact on CHD risk. Low-GI diets also appear to have little effect on lipid concentrations or HbA_{1c}, although most studies to date have been short term. Diets rich in protein or unsaturated fat have been found to reduce blood pressure more than diets high in carbohydrate in obese subjects with mild hypertension.

Ethanol

Several population studies have reported that the relationship between ethanol consumption and CHD risk is J-shaped, with a nadir for risk at approximately three units (30 mL) of ethanol per day. However, confounding factors make it difficult to interpret the relative benefits of moderate drinking. People who never drink

alcohol appear to have a higher prevalence of several other cardiovascular risk factors. Although systolic blood pressure rises with increasing alcohol intake (>3 units/day is associated with a significant increase in systolic blood pressure and other adverse events), alcohol intake is positively associated with HDL-cholesterol, particularly the HDL₂ subfraction. Plasma triglycerides also increase with alcohol consumption, but this is associated with an increase in VLDL and the large, buoyant LDL subfraction, rather than small, dense LDL. The relative benefits of various alcoholic beverages have been debated, and related to their antioxidant content. However, the form of alcohol consumed is likely to be confounded by lifestyle, diet and cultural factors. For example, the prevalence of wine drinking varies with age, race, smoking, ethnicity and educational background.

Fish and fish oils

A relationship between fish consumption and protection against CVD was first mooted in the 1960s because of the low rates of coronary disease observed in the Inuit people. Subsequent epidemiological studies have shown that consuming more than 35 g of fish per day is associated with a substantial reduction in CHD risk. This effect is dose dependent, each 20 g increase in fish intake being associated with a 7% lower risk for CHD mortality. The Nurses' Health Study population revealed similar findings with up to 34% lower CHD mortality when fish was eaten more than five times per week.

It has been proposed that the protection conferred by increasing fish consumption may relate to the original risk status of the population concerned, and the kind of fish consumed.

Soy protein

Soy is an edible protein component of soya bean. A meta-analysis of 38 clinical studies has shown that consumption of soy protein can lower plasma LDL-cholesterol by as much as 13% and triglycerides by 10%, but only in subjects with higher starting concentrations. Several mechanisms for this have been proposed but both soy proteins and isoflavones appear to be required. It is noteworthy that the per capita consumption in Japan (where the incidence of CHD is low) is as much as 55 g/day, whereas in the USA (where it is much higher), it is <5 g/day. An intake of >25 g/day is required to significantly improve the lipid profile.

Fatty acids

Changing the fatty acid composition of the diet can have a substantial effect on plasma cholesterol concentrations and on coronary risk. Dietary fatty acids can have a profound effect on the inflammatory process and consequently may also affect plaque composition and stability. A high intake of α -linolenic acid is associated with reduced risk of fatal CHD, but in men, an increased risk of prostate cancer has also been reported.

Plant sterols

Sterols are alcohol derivatives of cyclopentanoperhydrophenanthrene, and are essential constituents of all cell membranes. Cholesterol is a mammalian sterol, containing 27 carbon atoms, whereas sitosterol, campesterol and stigmasterol are plant sterols with 28 and 29 carbon atoms, due to the presence of an additional methyl or ethyl side chain, respectively. Dietary cholesterol has a modest effect on plasma cholesterol concentrations, and an intake of <300 mg/day is recommended for the general population (<200 mg/day in patients with CVD). Plant sterols are normally poorly absorbed by the human intestine and inhibit cholesterol absorption. Five g/day of sitostanol ester reduces plasma cholesterol concentration by approximately 10–20%.

Fibre

Dietary fibre intake is inversely related to CHD risk. A 10 g/day increase in fibre has been shown to be associated with a 14% decrease in coronary events and a 27% decrease in coronary deaths. Intake of whole grain wheat is associated with a 26% reduction in coronary disease, which may be related to a reduction in LDL and triglycerides. High dietary fibre intake is also associated with a reduced prevalence of diabetes, hypertension and obesity. It has been proposed that the fibre content of soy may account for some of its cholesterol-lowering properties. However, the effect of fibre on cholesterol is a modest one: there is approximately a 0.045 mmol/L fall in plasma cholesterol per gram of dietary fibre. The mechanisms of action are thought to include: bile salt binding, in much the same way that resins work; altered gut motility; an effect on satiety, and increased insulin sensitivity. Overall, high dietary fibre intake is associated with a substantially lower relative risk of fatal and non-fatal CVD. It should also be noted that soy contains phyto-oestrogens, and whole grain wheat is rich in selenium, so that the beneficial effects of these foods may not be entirely attributable to their fibre content.

Fruit and vegetables, tea and coffee

Higher rates of coronary disease have been found in regions with low fruit and vegetable consumption, and a diet rich in fruit and vegetables is associated with a 5% reduction in stroke risk per portion of fruit and vegetables per day. Vegetarians have been reported to have a lower rate of CHD, although these data are confounded by other associated patterns of lifestyle, including a lower fat and tobacco consumption and higher socioeconomic class. The Dietary Approaches to Stop Hypertension (DASH) trial found that a diet high in fruit and vegetables and low in fat reduced systolic blood pressure by 5.5 and diastolic blood pressure by 3 mmHg. Fruit and vegetables are also a rich source of antioxidants, such as vitamins C and E, carotene and polyphenols.

Green and black teas contain a complex mixture of polyphenolic compounds that include tannins. This spectrum of components and their relative content depends on the conditions of brewing and their bioavailability depends on whether the tea is taken with milk. However,

drinking tea more than three times per day is associated with reduced stroke and CHD. Drinking boiled coffee is associated with an increase in plasma cholesterol. Plasma homocysteine concentrations are also increased in subjects with a high coffee consumption. However, there is little evidence that coffee intake is related to overall cardiovascular risk.

Dietary pattern

While individual dietary constituents may ameliorate cardiovascular risk as outlined above, increasing emphasis is placed on having a healthy dietary pattern rather than individual dietary components. Healthy diet is one which, first, maintains BMI in the normal range and minimizes the risk of overweight or obesity. A number of scores of been developed to identify a healthy or prudent dietary pattern. However, one dietary pattern that has considerable evidence to support benefits for cardiovascular health and that is currently recommended in NICE cardiovascular disease prevention guidelines is the Mediterranean diet. The Lyon Heart Study examined the effects of a Mediterranean diet (30% of calories from fat, of which 8% were saturated) compared with a prudent Western diet (34% of calories from fat of which 12% was saturated fat) on coronary events in patients with established coronary disease. The Mediterranean diet contained more bread, fruit, root and green vegetables, poultry and fish, and less red meat; dairy products were replaced with a high α -linolenic acid margarine. After a 46-month follow-up period, there was a 50–70% lower risk of cardiovascular events in the Mediterranean diet group.

HYPERTENSION

Hypertension, or raised blood pressure, is one of the single most common and preventable causes of premature morbidity and mortality. It is a major risk factor for myocardial infarction; heart failure; left ventricular hypertrophy; stroke (both ischaemic and haemorrhagic); chronic kidney disease; peripheral vascular disease; cognitive decline, and premature death. It can also (rarely) present as malignant hypertension, an acute, life-threatening emergency (see below).

The laboratory is involved in:

- assessment of end-organ damage in all hypertensive patients
- monitoring for complications of some commonly used antihypertensive drugs
- investigation for causes of secondary hypertension, particularly screening for renal disease (all) and investigation of endocrine causes (a selected sub-group).

Definition

Blood pressure is measured as diastolic and systolic pressures, measured in millimetres of mercury (mmHg). Systolic pressure represents the peak blood pressure during ventricular contraction (systole); diastolic pressure represents the pressure during ventricular relaxation (diastole).

The blood pressure of individuals within a population is continuously distributed in a Gaussian (normal) manner. Although it would be possible to define a reference interval or range that comprised 95% of an apparently healthy population, this range (as with that for plasma cholesterol concentration) would include significant numbers of people whose blood pressure would cause increased risk of the long-term complications. There is no cut-off value above which hypertension definitively exists; thus any definition of hypertension is to some extent arbitrary.

Epidemiological studies demonstrate increased risk of adverse effects above blood pressures of 115/70 mmHg in all age groups. The relationship is continuous and progressive: for every 2 mmHg rise in blood pressure, the risk of death due to IHD increases by 7% and the risk of death from stroke by 10%. However, this observed association does not prove causality, which would require randomized trials demonstrating risk reduction with blood pressure reduction. Thus, the level of blood pressure determining the presence of hypertension is defined as the level of blood pressure above which treatment has been shown to reduce the development or progression of disease. This is currently generally accepted by most bodies internationally to be 140/90 mmHg.

Blood pressure measurement outside a clinic environment has been shown to correlate more strongly with hypertension-related morbidity and mortality than that measured in the clinic setting, and the latest (2011) NICE guidance recommends that home ambulatory blood

pressure monitoring be used to confirm hypertension detected in the clinic.

Hypertension is usually asymptomatic, and diagnosed incidentally, either as part of routine assessment in primary care, as part of a general medical examination when a patient presents with an unrelated condition or as a result of one of its complications (e.g. angina due to coronary atherosclerosis, breathlessness due to cardiac failure). Presenting features in malignant hypertension (see below) include headache, visual disturbances and fits.

Cause

In at least 90% of individuals with sustained hypertension, no primary cause is identifiable, and the condition is termed 'primary hypertension' (previously 'essential hypertension'). Up to 10% have 'secondary hypertension', for which it is possible to identify a specific cause, such as renovascular disease, phaeochromocytoma or hyperaldosteronism. The clinical picture can help target investigation appropriately (Table 38.2).

Primary hypertension

Primary hypertension is very common in Western populations and its prevalence is strongly influenced by age as well as lifestyle factors. Blood pressure tends to rise with age (to a greater extent in men than in women), although, over the age of 70 years, diastolic blood pressure may decline. It is estimated that at least 25% of the UK adult

TABLE 38.2 Clinical pointers to secondary hypertension

Clinical feature	Condition
Drug-resistant hypertension	Any
Acute hypertension with previously stable values	
Diagnosis before age 30 in non-obese	
Severe hypertension with evidence of end-organ damage	
Raised creatinine/reduced eGFR	Renal impairment
Acute rise in serum creatinine (>30%) with ACE inhibitor or angiotensin II receptor blockade	Renovascular hypertension
Moderate to severe hypertension with known vascular disease, asymmetry on renal ultrasound, flash pulmonary oedema	
Abdominal bruit (low sensitivity)	
Hypokalaemia (mild or easily-induced)	Primary aldosteronism (serum potassium more often normal)
Mild hypernatraemia	Primary aldosteronism
Cushingoid facies, truncal obesity, proximal myopathy, bruising	Cushing syndrome
Adrenal 'incidentaloma'	Primary aldosteronism
	Cushing disease
	Drug-induced
Oral contraceptives	
Glucocorticoids	
Ciclosporin	
Erythropoietin	
Illicit drug use (esp. cocaine, ecstasy)	
Excessive alcohol consumption	
Paroxysmal hypertension, esp. with headaches, sweating, palpitations	Phaeochromocytoma (rarely asymptomatic)
Obesity (particularly men)	Sleep apnoea syndrome
Heavy snoring	
Daytime drowsiness / headache	
Delayed or reduced femoral pulse	Coarctation of aorta
Aortic 'machinery' murmur	
Symptoms of hypothyroidism	Primary hypothyroidism
Raised TSH	
Hypercalcaemia	Primary hyperparathyroidism

population have hypertension, as well as more than 50% of those aged more than 60 years. Systolic pressure is more commonly elevated in patients in older patients, while raised diastolic pressure is a more prevalent feature of hypertension in those aged less than 50 years.

The aetiology of primary hypertension is poorly understood but is likely to be complex and multifactorial, including genetic, intrauterine and lifestyle factors. Increased sympathetic neural activity and increased angiotensin II activity have both been implicated, although no abnormalities have consistently been identified.

Blood pressure tends to exhibit a familial concordance, to an extent that cannot be explained by a shared environment: having one or more hypertensive parent approximately doubles an individual's risk of hypertension. An influence of intrauterine environment is indicated by the fact that babies born at term with lower birth weights tend to have higher blood pressure as adults, possibly as a result of reduced nephron mass.

Important lifestyle factors include obesity, high sodium intake and a high alcohol intake (though individuals who consume moderate alcohol tend to have slightly lower blood pressures than abstainers). A meta-analysis of 18 studies has recently indicated an association between hypertension and vitamin D deficiency.

The 'metabolic syndrome' comprises hyperinsulinaemia, glucose intolerance, central obesity and dyslipidaemia (all secondary to insulin resistance) and hypertension; insulin resistance and hypertension also frequently co-exist in patients with polycystic ovary syndrome, but in neither case is the mechanism linking insulin resistance and hypertension fully understood. Dyslipidaemia also shows an association with higher blood pressure, which is independent of body mass. Blood pressure also varies between populations, tending to be higher in black Africans than in white Europeans.

The prevalence of primary hypertension will continue to rise with increasing longevity, particularly with the concurrently rising prevalence of obesity. Lifestyle modification is an important aspect of the management of hypertension.

Secondary hypertension

Kidney disease. Hypertension is an important complication of both acute and chronic kidney disease, most commonly glomerular or vascular. It is detected by blood testing (elevated serum creatinine and urea) and/or dipstick urinalysis which may show proteinuria or haematuria. Disease is usually bilateral. Both activation of the renin-angiotensin-aldosterone axis and retention of sodium and water may contribute, the latter becoming more important with deteriorating renal function, and peripheral oedema may be observed. Less frequently, unilateral renovascular disease is responsible.

Endocrine disease. Pheochromocytoma, hyperaldosteronism and Cushing syndrome are all well-recognized (though individually relatively uncommon) causes of hypertension. Patients with thyrotoxicosis and hypothyroid patients may both have hypertension. There is also an association between primary hyperparathyroidism and

hypertension. Hypertension is present in some 40% of patients with acromegaly, although some of this is likely to be essential hypertension. In addition, there are several rare, inherited endocrine conditions whose manifestations include hypertension: these are summarized in [Table 38.3](#).

Laboratory assessment of hypertension

All patients should have formal cardiovascular risk prediction using one of the risk prediction equations discussed above. Electrolytes and creatinine (with estimation of glomerular filtration rate) should be measured as baseline before treatment is started with drugs which are potentially nephrotoxic or associated with deranged electrolytes. Renal damage can also both cause hypertension and be affected by target organ damage, and all newly diagnosed hypertensive patients should also be tested for haematuria using a reagent strip at the clinic, and for microalbuminuria with a urinary albumin:creatinine ratio.

Investigation for secondary causes

Comprehensive investigation of every patient with newly diagnosed hypertension is not an efficient use of health-care resources, since approximately 90% of patients will have primary hypertension and by definition no cause will be found. It is important to bear in mind and target patients with signs and symptoms suggesting a secondary cause of hypertension for specialist endocrine or radiological investigations. Specific clinical features that may suggest specific diagnoses are given in [Table 38.2](#). General features which should prompt comprehensive investigation include:

- severe or resistant hypertension. (Resistant hypertension is the persistence of hypertension despite concurrent use of adequate doses of three antihypertensive agents from different classes, including a diuretic. It is important to try to ascertain compliance, which is a common cause of apparently resistant hypertension.)
- an acute rise in blood pressure in a patient with previously well-controlled measurements
- aged less than 40 years in a Caucasian patient without a family history of hypertension or other risk factor such as obesity
- prepubertal onset
- malignant or accelerated hypertension (e.g. severe hypertension plus signs of end-organ damage such as retinal hemorrhages or papilloedema, heart failure or acute kidney injury).

Renovascular hypertension

A proportion of the 10% of all hypertensive patients who have secondary hypertension present with chronic kidney disease which, if not clinically apparent, will be obvious on initial laboratory screening. Renovascular hypertension is the most common remediable cause of secondary hypertension. It may occur in the setting of a patient with other known vascular disease such as CHD or peripheral vascular disease. Particularly suggestive clinical clues include:

TABLE 38.3 Inherited disorders causing hypertension

Condition	Inheritance	Plasma renin	Plasma aldosterone	Cause	Other features	Treatment
Glucocorticoid suppressible hyperaldosteronism	Autosomal dominant	Normal/↓	↓	Mutation causes fusing of regulatory sequence of steroid 11 β -hydroxylase with aldosterone synthase: aldosterone synthesis is controlled by ACTH rather than angiotensin II	Variable hypokalaemic alkalosis; ↑ plasma 18-hydroxycortisol	Dexamethasone
Liddle syndrome	Autosomal dominant	↓	↓	Activating mutation in distal renal tubular sodium transporter leads to sodium retention and ↑ ECF volume	Variable hypokalaemic alkalosis	Amiloride, triamterene
Gordon syndrome (pseudohypoaldosteronism type II)	Autosomal recessive	↓	↓	Mutation leading to increased distal renal tubular chloride reabsorption; hence increased sodium reabsorption and ↑ extracellular fluid volume	Hyperkalaemia, renal tubular acidosis	Thiazide diuretics; dietary salt restriction
Syndrome of apparent mineralocorticoid excess	Autosomal recessive	↓	↓	11 β -Hydroxysteroid dehydrogenase ^a deficiency, causing decreased metabolism of cortisol to cortisone; allows cortisol to act as a major mineralocorticoid, causing sodium retention	Hypokalaemic acidosis	Dexamethasone or high-dose spironolactone
Steroid 11 β -hydroxylase deficiency	Autosomal recessive	↓	↓	Enzyme deficiency leads to ↑ plasma 11-deoxycortisol and 11-deoxycorticosterone; the latter has mineralocorticoid activity	Hypokalaemic alkalosis; masculinization in females; premature virilization in males	Cortisol
Steroid 17 α -hydroxylase deficiency	Autosomal recessive	↓	↓	Enzyme deficiency leads to ↑ plasma 11-deoxycorticosterone and corticosterone; the former has mineralocorticoid activity	Hypokalaemic alkalosis; feminization in males; lack of normal sexual maturation in females	Cortisol

^aThis enzyme converts cortisol to cortisone (inactive) but does not metabolize aldosterone; its activity normally limits the mineralocorticoid activity of cortisol. Carbenoxolone (glycyrrhetic acid) and liquorice (which contains glycyrrhizic acid) are inhibitors of this enzyme and can cause hypertension and hypokalaemia.

- the development of an acute sustained rise in serum creatinine (>50%) within a week of starting treatment with an angiotensin-converting-enzyme inhibitor or angiotensin II receptor blocker
- severe hypertension in a patient with asymmetric kidneys or an unexplained atrophic kidney seen on renal ultrasound
- severe hypertension in a patient with recurrent acute pulmonary oedema, or severe heart failure with impaired renal function
- severe hypertension in conjunction with a systolic–diastolic abdominal bruit that lateralizes to one side.

Any of these features should trigger prompt discussion with the local nephrology service and imaging of the renal arteries.

Primary aldosteronism (hyperaldosteronism)

Primary aldosteronism (PA) is the commonest identifiable, specifically treatable and potentially curable form of hypertension. The triad of hypertension, unexplained hypokalaemia, and an aldosterone-producing adenoma (APA) of the adrenal gland, was first described in 1955 by Jerome Conn, in a 34-year-old female who had a seven-year history of intermittent muscle weakness, muscle

spasms and cramping of her hands. This clinical picture is not specific for Conn adenoma but is associated with any cause of mineralocorticoid excess. Metabolic alkalosis is also a common feature.

Primary aldosteronism comprises a group of disorders characterized by inappropriately high secretion of aldosterone, which is autonomous and is not suppressed by saline. A growing body of research has implicated aldosterone in the pathogenesis of vascular disease independently of blood pressure, possibly via promotion of systemic inflammation and oxidative stress, which contribute to insulin resistance and vascular abnormalities. The recognition that most patients with PA (>75%) are actually normokalaemic has led to an appreciation that it is much more common than previously recognized. As many as 5–13% of patients with hypertension are currently thought to have PA.

Forms of PA

Around 30% have a unilateral APA, which is potentially resectable by laparoscopic adrenalectomy (hypokalaemia is more common in this group). Most of the remainder have bilateral idiopathic hyperaldosteronism (IHA), which responds best to aldosterone-antagonist drugs. A rare cause is glucocorticoid-suppressible aldosteronism (GRA). Very rarely, an adrenal carcinoma is responsible, and it is occasionally a component of multiple endocrine neoplasia (MEN) type 1. In general, hypertension and hypokalaemia are more severe in patients with APA than bilateral forms.

The diagnosis of PA should be considered in:

- hypertensive patients with spontaneous or diuretic-induced hypokalaemia (see Table 38.4 for differential diagnoses)

TABLE 38.4 Differential diagnosis of primary aldosteronism (PA)

Hypertensive condition	Differentiating clinical features	Differentiating tests
Primary hypertension	None	Aldosterone/renin ratio normal <i>off interferent drugs</i> × 2–4 weeks. Normal ratio on drugs that can potentially cause false positives makes PA highly unlikely Hypokalaemia not a feature of primary HT (only infrequently seen in PA)
Diuretic-induced hypokalaemia in patient with primary hypertension	Drug history	Aldosterone/renin ratio normal after correction of hypokalaemia and diuretic withdrawal for at least six weeks
Secondary hypertension	Known vascular disease or risk factors; known renal artery stenosis	Aldosterone/renin ratio normal or low Imaging studies (to demonstrate renal artery stenosis or reninoma)
Ectopic ACTH syndrome	Features of underlying malignancy (usually bronchial small cell tumour; rarely thymus pancreas or bronchial carcinoid) Rarely Cushingoid features (may be present if slow-growing tumour)	Aldosterone/renin ratio usually normal despite hypokalaemia and suppressed renin (aldosterone also low) Evidence of malignancy on imaging studies Cortisol and ACTH levels elevated and non-suppressible with high-dose dexamethasone
Liddle syndrome	Childhood presentation. Family history (autosomal dominant)	Aldosterone/renin ratio usually normal despite hypokalaemia and suppressed renin (aldosterone also low)
Syndrome of apparent mineralocorticoid excess	Childhood presentation of primary form. Family history (autosomal recessive) May be acquired due to excessive liquorice consumption	Aldosterone/renin ratio usually normal despite hypokalaemia and suppressed renin (aldosterone also low) Raised urinary free cortisol/cortisone ratio
Hypertensive forms of congenital adrenal hyperplasia	Childhood presentation. History of either virilization (in 11 β -hydroxylase deficiency) or feminization (in 17 α -hydroxylase deficiency) Family history of congenital adrenal hyperplasia due to 11 β -hydroxylase or 17 α -hydroxylase deficiency (autosomal recessive)	Aldosterone/renin ratio usually normal despite hypokalaemia and suppressed renin (aldosterone also low) 11 β -hydroxylase deficiency: low plasma cortisol and corticosterone; raised basal or ACTH-stimulated levels of deoxycorticosterone and 11-deoxycortisol 17 α -hydroxylase deficiency: low plasma 17 α -hydroxyprogesterone, 11-deoxycortisol, and cortisol; increased gonadotrophins (LH and FSH)
Primary glucocorticoid resistance	Family history (although may be acquired) May be associated with androgenization	Aldosterone/renin ratio normal despite hypokalaemia and suppressed renin (aldosterone also low) Raised ACTH and cortisol; resistance of cortisol to dexamethasone suppression in absence of clinical features of Cushing syndrome
Activating mutations of the mineralocorticoid receptor	Family history Pregnancy-induced worsening of HT and development of hypokalaemia	Aldosterone/renin ratio usually normal despite hypokalaemia and suppressed renin (aldosterone also low)
Pseudohypoaldosteronism type II (Gordon syndrome)	Family history (autosomal dominant) Hyperkalaemia; mild normal anion gap metabolic acidosis Increased distal tubular Cl ⁻ reabsorption	Low renin, normal aldosterone; often raised aldosterone/renin ratio

- patients with moderate, severe or resistant hypertension (less common in mild hypertension)
- patients with adrenal incidentaloma and hypertension
- patients with the onset of hypertension under the age of 20
- patients with hypertension and a family history of young-onset hypertension or CVA at young age (<40 years), who may have the rare, inherited glucocorticoid-remediable form of PA.

Biochemical investigation

Determination of the ratio of plasma aldosterone concentration to plasma renin activity is widely accepted as the optimal test for screening, after controlling for factors (including medicines) that may confound results. (See Appendix 1 for detailed protocol.)

Confounding factors

Posture. Assuming upright posture effects a rise in plasma aldosterone via increased renin, released from the juxtaglomerular apparatus in response to a slight fall in renal perfusion. All APA, and most with IHA, retain normal responsiveness to upright posture, and measurement in a sample taken when upright may be the most sensitive in detecting raised aldosterone. In practice, a mid-morning upright sample, taken after sitting for 5–15 min appears to be sensitive as well as pragmatic.

Time of day. In PA patients, suppression of renin makes aldosterone more strongly influenced by ACTH, which has a circadian rhythm, being highest in the morning, then falling rapidly. Thus, morning samples are most likely to show elevated aldosterone.

Drugs. See Box 38.4.

Relevant drugs should be withdrawn at least two weeks before testing, or for diuretics, at least four weeks. Where antihypertensive drugs cannot be safely withheld, verapamil slow-release, hydralazine and/or prazosin have lesser effects on the aldosterone–renin system.

BOX 38.4 Drugs known to potentially interfere with aldosterone:renin ratio

Drugs potentially causing false positive aldosterone:renin ratio

- β -Adrenergic receptor blockers
- Methyldopa
- Clonidine
- Non-steroidal anti-inflammatory drugs
- Renin inhibitors (if plasma renin activity measured)
- Oestrogen-containing drugs (if direct active renin measured)

Drugs potentially causing false negative aldosterone renin ratio

- Diuretics (all classes)
- Dihydropyridine Ca-channel antagonists
- Angiotensin-converting-enzyme (ACE) inhibitors
- AII receptor blockers (ARBs)
- Renin inhibitors (if direct active renin measured)

Dietary sodium. Salt restriction stimulates renin production, which may lower the aldosterone:renin ratio, thus patients are advised to maintain a liberal salt intake before testing.

Plasma potassium. Hypokalaemia suppresses aldosterone secretion and may cause a false negative result, thus any hypokalaemia should be corrected as far as possible with oral potassium supplements for several days prior to testing.

Phase of menstrual cycle. Higher aldosterone concentrations seem to be measured in the luteal phase relative to the follicular phase, almost certainly mediated by sex hormones, and it may thus be preferable to screen premenopausal women early in the menstrual cycle. More research is required in this area before firm recommendations can be given.

Renal impairment/elderly patients. Renin levels are lower in renal impairment due to reduced nephron mass as well as sodium and water retention. Similarly elderly patients secrete less renin and false positive ratios may result in both circumstances.

Other conditions. Co-existing conditions, which may give rise to false negatives, include pregnancy, renal artery stenosis and malignant hypertension.

Differences between laboratories in assays used and units reported give rise to variability in cut-off values used by different groups; this makes standardization of diagnostic criteria difficult. Renin measurement can be in terms of its enzymatic activity (plasma renin activity, PRA) or its mass (direct active renin concentration, DAR). Automated immunometric assays for measuring DAR have been widely adopted as they are faster and more convenient than PRA. However, considerable work is required to validate these methods, and in the meantime, PRA remains the preferred measurement.

Because of inherent variability in both renin and aldosterone, measurements should be repeated at least once for confirmation. In patients with repeatedly elevated aldosterone/renin ratios, definitive confirmation or exclusion of diagnosis involves careful suppression testing with measurement of aldosterone response to fludrocortisone or to salt loading.

Genetic testing for the hybrid gene causing familial hyperaldosteronism type I (glucocorticoid-remediable aldosteronism (GRA)) allows subtype differentiation, which allows treatment to be tailored. However this is rare, and the majority of genetic tests will be negative.

Localization

The next step should be to differentiate unilateral APA amenable to surgery from bilateral forms of PA which respond to drugs. Computerized tomography (CT) and MRI scanning are both used. Adrenal CT on its own is unreliable, as it may miss small secreting APAs, and misdiagnose larger non-functional adrenal tumours ('incidentalomas') as APA, leading to unnecessary surgery. The most sensitive technique for the localization of an isolated adenoma is selective adrenal vein blood sampling with measurement of the aldosterone/cortisol ratio, although this is a technically demanding technique.

The investigation of Cushing syndrome is discussed in Chapter 18.

Phaeochromocytoma

Phaeochromocytoma and functional paraganglioma (PGL) are catecholamine-secreting tumours of the adrenal medulla and extra-adrenal sympathetic nervous tissue (e.g. sympathetic ganglia), respectively. Because of their similarity in presentation and management principles, the term phaeochromocytoma is often used to referred to both, and this convention is also used here. Approximately 10% of phaeochromocytomas are extra-adrenal.

A patient identified with phaeochromocytoma is usually symptomatic. The classical triad of paroxysmal headache, sweating and palpitations is due to excess tumoral secretion of noradrenaline, adrenalin, and dopamine, and is said to have high predictive value. Pallor, lethargy, tremor, nausea, flushing and orthostatic hypotension can also occur. Episodes last from minutes up to an hour and are self-limiting. Symptoms are characteristically made worse by β -blockade because their use permits unopposed α -adrenergic stimulation.

Prevalence in the general population is uncertain, but it accounts for fewer than 0.2% of patients with hypertension. Its importance lies in its being malignant in about 10% of patients (more frequently in the context of multiple endocrine neoplasia, MEN, type 2) and being potentially curable. Malignant tumours biochemically and histologically resemble benign tumours, and can only be diagnosed for certain by the presence of local or distant spread which may occur many years after the primary tumour is resected.

Approximately 90% of phaeochromocytomas are sporadic: these are usually single adenomas. Approximately 10% are inherited: these are often bilateral and multiple. Inherited phaeochromocytomas may be isolated (inherited as an autosomal dominant trait) or be a component of multiple endocrine neoplasia (MEN 2A and 2B, in which they have a combined prevalence of 40%). They also occur as the sole manifestation of MEN 2A in approximately 25% of heterozygotes for this condition. (See Chapter 41 for further details of inherited phaeochromocytomas.)

Patients with phaeochromocytomas may come to medical attention because of their hypertension (paroxysmal in about one-third of patients, but more often sustained though variable) or because of the clinical features of increased secretion of catecholamines.

The diagnosis should be suspected in individuals with one or more of the following:

- paroxysmal self-limiting episodes of typical adrenergic symptoms (see above)
- resistant or labile hypertension (although 5–15% are normotensive at diagnosis)
- familial syndrome predisposing to catecholamine-secreting tumours (e.g. MEN2)
- family history of phaeochromocytoma
- incidentally discovered adrenal mass (25% of diagnoses)
- onset of hypertension at a young age (e.g. <20 years)
- hypertension with new onset or atypical diabetes mellitus
- hypertensive response during anesthesia, surgery or angiography

- idiopathic dilated cardiomyopathy
- history of gastric stromal tumour or pulmonary chondromas (Carney triad).

Biochemical investigation

Diagnosis of phaeochromocytoma involves first, demonstrating excessive secretion of catecholamines, followed by radiological localization of the causative tumour(s). Consideration of the diagnosis and a low threshold for investigation are key initial factors in successful diagnosis. Because the symptoms are so non-specific, diagnosis is rarely confirmed biochemically in patients with suggestive symptoms. The investigation strategy in a particular patient will be dictated by the availability of biochemical tests in any centre. However, as it is so important not to miss the diagnosis, testing with a high diagnostic sensitivity is critical, particularly when screening asymptomatic patients with an inherited risk of phaeochromocytoma.

Traditional measurement of urinary catecholamines is gradually being superseded by HPLC measurement of fractionated metanephrines (metabolites of catecholamines), either in a 24 h urine collection or in plasma, both of which offer better diagnostic sensitivity and specificity (Table 38.5). The improved sensitivity is a result of continuous production of metanephrines by the tumour, which contains a high concentration of catechol-O-methyltransferase. Even when the tumour release of catecholamines by the tumour is sporadic, this enzyme ensures continued metabolism of catecholamines to metanephrines. The sensitivity of metanephrine measurement to detect phaeochromocytoma is 99% for plasma and 97% for urine, which compare favourably with 86% for urinary catecholamines, among symptomatic patients, those with a previous phaeochromocytoma or a family history, and those with an incidental adrenal mass. The diagnostic specificities for plasma and urinary metanephrines are 89% and 93%, respectively, compared with 88% for urinary catecholamines. False positive results are common owing to the presence of other medical conditions (such as obstructive sleep apnoea) and medications (Table 38.6) that can cause elevation of metanephrines. The relative infrequency of phaeochromocytoma in those tested also contributes to the high rate of false positives. In addition,

TABLE 38.5 Sensitivities and specificities of biochemical tests for phaeochromocytoma in 214 patients

	Sensitivity (%)	Specificity (%)
P free metadrenalines	99	89
U fractionated metadrenalines (HPLC)	97	64
P catecholamines	81	81
U catecholamines	85	86
U total metadrenalines (spectrophotometry)	74	93
U vanillylmandelic acid (VMA)	62	93

Data from Lenders et al. JAMA 2002; 287:1427–1434. P, plasma; U, urine.

TABLE 38.6 Some drug interferences with measurement of catecholamines and metadrenalines

Interferent	Effect
Tricyclic antidepressants	↑ catecholamines and metadrenalines in U and P
Phenoxybenzamine	↑ catecholamines and metadrenalines in U and P
Monoamine oxidase inhibitors	↑ metadrenalines in U and P
Buspirone	↑ metadrenalines in U
Calcium-channel blockers	↑ catecholamines in U and P
Nicotine	↑ catecholamines in U and P
Caffeine	↑ catecholamines in U and P Analytical interference with HPLC assays
Methyldopa	Analytical interference with HPLC assays
Labetalol	Analytical interference with HPLC assays
Levodopa	Analytical interference with HPLC assays ↑ catecholamines and metadrenalines in U and P?

U, urine; P, plasma.

physical and psychological stress (e.g. surgery, severe pain or anxiety) can potentially increase catecholamine secretion, leading to false positive results.

Concentrations of metanephrines that are over four times normal are associated with a very high probability of pheochromocytoma; however, intermediate raised levels will require further biochemical investigation, which may include repeat testing after withdrawal of implicated medication. By contrast with the, now infrequently used, measurement of urine vanillylmandelic acid, dietary restrictions are not needed when performing these investigations.

Localization

The purpose of localization is to guide surgical intervention after biochemical confirmation, as well as to look for possible metastatic disease. Direct anatomical imaging methods by computed (CT) tomography or magnetic resonance imaging (MRI) should be used as first-line to locate the primary tumours and is particularly helpful in the relatively large, sporadic, symptomatic tumours. Both techniques have high sensitivity but lower specificity, due to the high prevalence of the benign adrenal cortical adenoma ('adrenal incidentaloma'). Meta-iodobenzylguanidine (MIBG) scintigraphy is particularly helpful for detecting smaller, presymptomatic tumours, extra-adrenal tumours, including metastases and recurrent tumours. However, false positive results may occur in patients with carcinoid tumours and, because the isotopes are excreted in the urine, false negatives can occur with tumours in the bladder or head of the pancreas (which is closely related to the right kidney). Overall, MIBG scanning has higher specificity than anatomical imaging. Newer functional imaging techniques include ¹⁸F-fluorodopamine positron

emission tomography (PET) and ¹¹¹In-octreotide single photon emission computed tomography (SPECT) scanning.

Management

It is beyond the scope of this book to discuss the management of pheochromocytoma in any detail. Even following apparently successful surgical resection, patients should be followed-up long term, both clinically (blood pressure measurements) and biochemically. Hypertension may persist and tumours may recur.

Malignant hypertension

Malignant hypertension is defined as markedly elevated blood pressure (diastolic BP often >140mmHg) with retinal haemorrhages, exudates or papilloedema with or without evidence of target organ damage, typically acute and progressive kidney injury with proteinuria and haematuria ('malignant nephrosclerosis'). Its management is a medical emergency and should take precedence over investigations to determine whether there is an underlying cause. Neurological signs in malignant hypertension may be due to intracerebral or subarachnoid haemorrhage or lacunar infarct, or to hypertensive encephalopathy. The latter is associated with insidious onset and non-lateralizing symptoms, such as headache, vomiting, restlessness, confusion. It may be confirmed radiologically on the basis of MRI, which shows oedema of the white matter in the occipito-parietal regions, termed 'reversible posterior leucoencephalopathy'. (The distinction is important as hypertensive encephalopathy mandates aggressive reduction of blood pressure, not usually indicated for haemorrhage or infarct.)

Hypertension in pregnancy

Hypertension developing in pregnancy constitutes a significant risk to the mother and the fetus. Pre-eclampsia is the development of new hypertension and proteinuria in the second half of pregnancy. It is a multisystem disorder and if untreated, the condition can progress to severe hypertension, acute kidney injury and seizures (eclampsia) and with high fetal and maternal morbidity and mortality. An increase in plasma urate concentration can be a sensitive indicator of deteriorating renal function in this condition. Low platelet count, microangiopathic haemolysis seen on blood film examination, rising creatinine and liver transaminases appear to be particularly predictive of adverse outcome. Pre-eclampsia may sometimes be a feature of the haemolysis, elevated liver enzymes and low platelet (HELLP) syndrome, which is discussed in Chapter 22.

Management of hypertension

Except in malignant hypertension, or in secondary causes of hypertension for which there are specific treatments, management should always begin with efforts to bring about appropriate lifestyle changes, for example attainment of ideal body weight, reduction in any excessive

alcohol intake, reduction in dietary salt intake and attention to other cardiovascular risk factors.

The classes of drugs available include diuretics, calcium channel antagonists, α - and β -blockers, ACE inhibitors and angiotensin receptor antagonists, and centrally acting agents. Combinations of drugs are often required to achieve adequate control of blood pressure; if three or more are required in a young (<40 years) patient with normal renal function, a secondary cause of hypertension should be considered. Compliance with treatment can pose a considerable challenge, particularly because hypertension is generally asymptomatic but also because drug treatment may cause significant adverse effects.

CONCLUSION

Biochemical investigations play a key role in the management of several distinct areas of cardiovascular disease.

Plasma concentrations of markers of myocyte necrosis form the basis of evidence-based criteria that influence decisions in the management of acute coronary syndrome. Increase in cardiac troponin concentrations is embedded in the universal definition of myocardial infarction, and new cardiac biomarkers, as well as panels of biomarkers with complementary timeframes, are currently being developed and evaluated.

The diagnostic and prognostic value of natriuretic peptides in heart failure is reflected in their inclusion as recommended first-line investigations, although they have limitations including lack of specificity.

Measurement of plasma lipids is well-established as a key component of the estimation of cardiovascular risk, and work is currently underway to examine the potential role of inflammatory markers in further stratifying this risk.

The biochemistry laboratory is involved in investigation of end-organ damage in hypertension, monitoring for side-effects of antihypertensive agents, and in the diagnosis of causes of secondary hypertension. A low threshold for investigation is necessary to avoid missing disorders such as pheochromocytoma and primary aldosteronism, which are probably underdiagnosed.

Cardiovascular disorders are associated with high morbidity and mortality. They contribute significantly to the workload of modern routine clinical biochemistry laboratories in developed countries, and are likely to continue to do so for the foreseeable future.

Further reading

Atherogenesis

Goldstein JL, Ho YK, Basu SK et al. Binding site on macrophages that mediates uptake and degradation of acetylated low density lipoprotein, producing massive cholesterol deposition. *Proc Natl Acad USA* 1979;76:333–7.

Key paper providing the first comprehensive description of the scavenger receptor.

Wong BW, Meredith A, Lin D et al. The biological role of inflammation in atherosclerosis. *Can J Cardiol* 2012;28:631–41.

A review of the current understanding of atherogenesis with a particular focus on the role of inflammation.

Myocardial injury

Gaze DC, Collinson PO. Multiple molecular forms of circulating cardiac troponin: analytical and clinical significance. *Ann Clin Biochem* 2008;45:349–55.

An interesting description of the biochemical structure of troponin molecules, including discussion of the significance of post-translational modifications.

Kelley WE, Januzzi JL, Christenson RH. Increases of cardiac troponin in conditions other than acute coronary syndrome and heart failure. *Clin Chem* 2009;55:2098–112.

A well-written and wide-ranging review of reports of non-cardiac troponin elevations.

McCann CJ, Glover BM, Menown IB et al. Novel biomarkers in early diagnosis of acute myocardial infarction compared with cardiac troponin. *Eur Heart J* 2008;29:2843–50.

An indication of what may succeed troponins as cardiac biomarkers of the future, with particular emphasis on H-FABP.

Thygesen K, Alpert JS, Jaffe AS et al. Joint European Society of Cardiology/American College of Cardiology Foundation/American Heart Association/World Heart Federation Task Force for the Universal Definition of Myocardial Infarction. Third universal definition of myocardial infarction. *Eur Heart J* 2012;33:2551–67.

Updated version of the 2007 universal guideline, which embeds laboratory measurement of troponin as the basis of diagnosis of MI. Claimed with some justification to be a 'global document'.

Heart failure

van Kimmenade RR, Januzzi Jr. JL. Emerging biomarkers in heart failure. *Clin Chem* 2012;58:127–38.

A good review of established and emerging biomarkers in heart failure.

Cardiovascular risk factors and prevention

Fifth Joint Task Force of the European Society of Cardiology, European Association of Echocardiography, European Association of Percutaneous Cardiovascular Interventions et al. European Guidelines on cardiovascular disease prevention in clinical practice (version 2012): the Fifth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of nine societies and by invited experts). *European Journal of Preventive Cardiology* 2012;19:585–667.

Hypertension

Peaston RT, Ball S. Biochemical detection of pheochromocytoma: why are we continuing to ignore the evidence? *Ann Clin Biochem* 2008;45:6–10.

A thought-provoking paper supporting a low threshold for screening with metadrenalines

Stowasser M, Taylor PJ, Pimenta E et al. Laboratory investigation of primary aldosteronism. *Clin Biochem Rev* 2010;31:39–56.

A comprehensive account of how, who and when to screen for this group of disorders, with good discussion of potential pitfalls.

APPENDIX 1: PROTOCOL FOR INVESTIGATION OF ALDOSTERONISM: SCREENING AND CONFIRMATORY TESTS

Patient preparation

Ideally, all drugs that have an effect on the renin–aldosterone system should be discontinued for two weeks before samples are collected. Aldosterone antagonists (e.g. spironolactone) and oestrogens must be discontinued for at least six weeks. Where antihypertensive drugs cannot safely be withheld, verapamil (slow-release), hydralazine, doxazosin or prazosin can be used as these have minimal effect on the renin–aldosterone system. If essential to continue other drugs to control blood pressure, as a minimum, β -blockers and ACE inhibitors should be stopped for two weeks and calcium channel blockers withheld on the day of the test until after its completion. The potential effects of interfering drugs must be taken into account in the interpretation of the renin and aldosterone results if they have to be continued (see Box 38.4).

The patient must be encouraged to take a liberal amount of dietary sodium (typically 100–150 mmol/day). Administer potassium salts orally, or if necessary intravenously, to restore plasma potassium concentration to within the reference range or, if this is not attainable, to the maximum concentration possible. Discontinue this supplementation 24 h before blood samples are taken. The patient must be well hydrated at the time of the test.

Screening procedure

Collect blood mid-morning, after the patient has been up (sitting, standing or walking) for at least two hours and then seated for 5–15 min.

Draw blood into a tube containing anticoagulant (type required is assay dependent). The sample should be taken rapidly to the laboratory, but **not** on ice as cooling of the sample promotes conversion of inactive renin precursor to active renin. Care must be taken to avoid haemolysis, which will result in an artefactual increase in plasma potassium concentration.

Measure plasma aldosterone, renin and electrolytes.

Interpretation of results of screening test

A high aldosterone:renin ratio indicates probable primary aldosteronism (PA). The particular decision limits used are method dependent and therefore vary from laboratory to laboratory, although if the ratio of aldosterone (in pmol/L) to PRA (in pmol/mL/h) is >2000 , the patient almost certainly has PA. Patients with a positive screening test should proceed to a confirmatory test before further investigations are undertaken to determine the cause and location of the disease. If either aldosterone or renin results are abnormal but the ratio is not clearly diagnostic, the screening test should be repeated after a further period off-treatment from all potentially interfering drugs, if some had initially been continued. Other causes of abnormal aldosterone or renin results should be reviewed (see Table 38.4).

CONFIRMATORY TESTS

A number of different confirmatory tests have been described, and there is no definitive evidence as to which has the best diagnostic performance. Of the two tests most commonly used in the UK, the saline suppression test is more convenient and safer to perform than the fludrocortisone suppression test.

Saline suppression test

The test is only indicated when the screening test demonstrates a high aldosterone:renin ratio. The same patient preparation requirements as for the screening test apply. The saline suppression test is contraindicated in patients with severe uncontrolled hypertension, chronic kidney disease, heart failure, cardiac arrhythmia or severe hypokalemia.

Procedure

1. Start test between 08.00 and 09.30 h.
2. Place indwelling catheter in antecubital fossa for infusion of 0.9% saline.

3. Place indwelling catheter in opposite arm for blood sampling.
4. Position patient in recumbent position for 1 h prior to commencing the infusion, and throughout test. Blood pressure and heart rate must be monitored throughout the test.
5. Draw blood into a tube containing anticoagulant (type required is assay dependent), for aldosterone, renin and electrolytes – do not place on ice but send to the laboratory immediately.
6. Infuse 2 L of 0.9% saline over 4 h.
7. At completion of infusion, immediately (with patient still recumbent) draw another blood sample for measurement of aldosterone and renin – do not place on ice but send to the laboratory immediately.

Interpretation

Plasma aldosterone concentration of >140 pmol/L at the end of the test confirms a diagnosis of PA.

Fludrocortisone suppression test

The test is only indicated when the screening test demonstrates a high aldosterone:renin ratio. The same patient preparation requirements as for the screening test apply. The fludrocortisone suppression test is contraindicated in the elderly and those with severe hypertension owing to the risks associated with sodium retention. This test should be performed on an in-patient basis.

Procedure

1. Day 1 (mid-morning). Ensure that the patient is upright for at least 30 min then draw blood into a tube containing anticoagulant (type required is assay dependent), for aldosterone, renin and electrolytes – do not place on ice but send to the laboratory immediately.
2. Administer fludrocortisone 0.1 mg 6-hourly for four days.
3. Administer oral slow-release sodium tablets 30 mmol 8-hourly with meals for four days.
4. Administer slow-release potassium tablets in sufficient quantity to maintain plasma potassium concentration close to 4.0 mmol/L. Monitor plasma potassium concentration at least twice daily.
5. Day 4 (mid-morning). Ensure that the patient is upright for at least 30 min and repeat blood sample as on day 1.

Interpretation

Upright plasma aldosterone >170 pmol/L on day 4 confirms PA, provided renin is fully suppressed.

Reference

Funder JW, Carey RM, Fardella C et al. Case Detection, Diagnosis, and Treatment of Patients with Primary Aldosteronism: An Endocrine Society Clinical Practice Guideline. *J Clin Endocrinol Metab* 2008;93:3266–3281.

Therapeutic drug monitoring

Mike Hallworth

CHAPTER OUTLINE

INTRODUCTION 767

Pharmacokinetics and pharmacodynamics 767
Which drugs should be measured? 769

USE OF THERAPEUTIC DRUG MONITORING 771

Appropriate clinical question 771
Accurate patient information 772
Appropriate sample 772
Accurate analysis 774
Relevant clinical interpretation 774
Effective action taken 774

PROVISION OF A THERAPEUTIC DRUG MONITORING SERVICE 775

Staff 775
Turnaround time 775
Point-of-care testing 775
Reporting 775
Units 775

Quality assurance 776
Continuing education 776

PHARMACODYNAMIC MONITORING, BIOMARKERS AND PHARMACOGENETICS 776

Integrating information 777

INDIVIDUAL DRUGS 777

Analgesic/anti-inflammatory drugs 777
Antiarrhythmics and cardiac glycosides 777
Anticonvulsants (antiepileptics) 778
Antidepressants and antipsychotic drugs 780
Antimicrobial drugs 781
Antineoplastic drugs 782
Bronchodilator drugs 783
Immunosuppressants 783
Opiate and opioid drugs 785

APPENDIX 786

INTRODUCTION

The aim of therapeutic drug monitoring (TDM) is to aid the clinician in the choice of drug dosage in order to provide the optimum treatment for the patient and, in particular, to avoid iatrogenic toxicity. It can be based on pharmacogenetic, demographic and clinical information alone (*a priori* TDM), but is normally supplemented with measurement of drug or metabolite concentrations in blood or markers of clinical effect (*a posteriori* TDM). Measurements of drug or metabolite concentrations are only useful where there is a known relationship between the plasma concentration and the clinical effect, no immediate simple clinical or other indication of effectiveness or toxicity and a defined concentration limit above which toxicity is likely. Therapeutic drug monitoring has an established place in enabling optimization of therapy in such cases.

Pharmacokinetics and pharmacodynamics

Before discussing which drugs to analyse or how to carry out the analyses, it is necessary to review the basic

elements of pharmacokinetics and pharmacodynamics. Essentially, pharmacokinetics may be defined as what the body does to drugs (the processes of absorption, distribution, metabolism and excretion), and pharmacodynamics as what drugs do to the body (the interaction of pharmacologically active substances with target sites (receptors) and the biochemical and physiological consequences of these interactions). The processes involved in drug handling are summarized in [Figure 39.1](#), which also indicates the relationship between pharmacokinetics and pharmacodynamics, and will now be discussed briefly. For a more mathematical treatment, the reader is referred to the pharmacokinetic texts listed in Further reading.

Adherence

The first requirement for a drug to exert a clinical effect is obviously for the patient to take it in accordance with the prescribed regimen. Patients are highly motivated to comply with medication in the acute stages of a painful or debilitating illness, but as they recover and the purpose of medication becomes prophylactic, it is

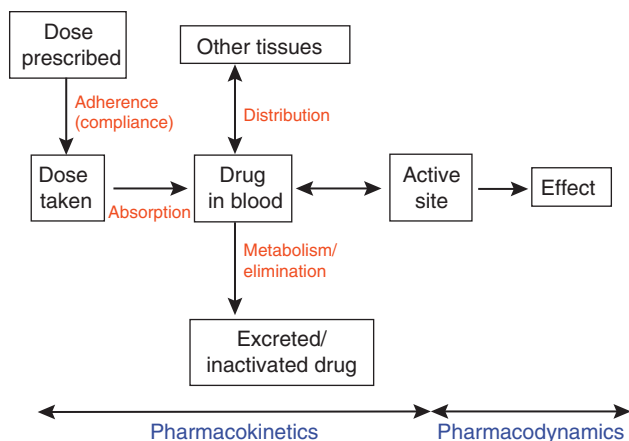


FIGURE 39.1 ■ Processes involved in drug handling.

easy for them to underestimate the importance of regular dosing. ‘Compliance’ is the traditional term used to describe whether the patient takes medication as prescribed, though ‘adherence’ has become the preferred term in recent years as being more consistent with the concept of a partnership between the patient and the clinician. ‘Compliance’, however, remains in widespread use. In chronic disease states such as asthma, epilepsy or bipolar disorder, variable adherence is widespread. The fact that a patient is not taking a drug at all is readily detectable by measurement of drug concentrations, although variable adherence may be more difficult to identify by monitoring plasma concentrations.

Absorption

Once a drug has been taken orally, it needs to be absorbed into the systemic circulation. This process is described by the pharmacokinetic parameter *bioavailability*, defined as the fraction of the absorbed dose that reaches the systemic circulation. Bioavailability varies between individuals, between drugs and between different dosage forms of the same drug. In the case of intravenous administration, all of the drug goes directly into the systemic circulation and bioavailability is 100% by definition, but different oral formulations of the same drug may have different bioavailability depending, for example on the particular salt or packing material that has been used. Changing the formulation used may require dosage adjustment guided by TDM to ensure that an individual’s exposure to drug remains constant. Other routes of absorption such as intramuscular, subcutaneous or sublingual may exhibit incomplete bioavailability. The total amount of drug absorbed can be determined from the area under the plasma concentration/time curve (see Fig. 39.2).

Distribution

When a drug reaches the bloodstream, the process of distribution to other compartments of the body begins. The extent of distribution of a drug is governed by its relative solubility in fat and water, and by the binding capabilities of the drug to plasma proteins and tissues. Drugs that are strongly bound to plasma proteins and exhibit

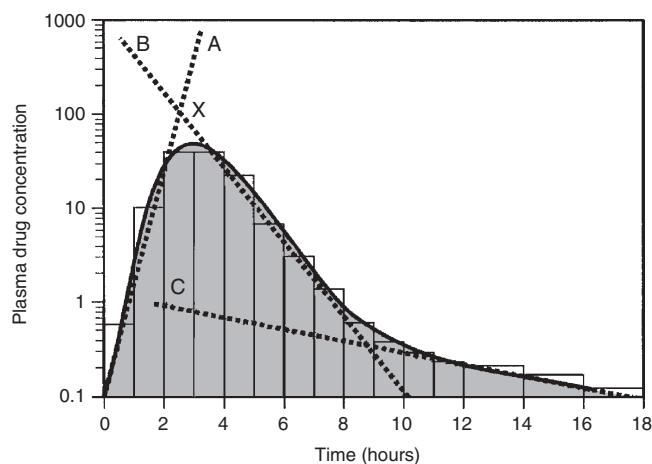


FIGURE 39.2 ■ Concentration/time curve for a drug administered by any route other than intravenously. Line A represents the absorption phase; line B indicates the distribution half-life and line C the elimination half-life. At point X, absorption, distribution and metabolism are all in progress. The areas of the rectangles can be summed to estimate the total amount of drug absorbed, the ‘area under the curve’ (AUC) (hatched area).

low lipid solubility and tissue binding will be retained in the plasma and show minimal distribution into tissue fluids. Conversely, high lipid solubility combined with low binding to plasma proteins will result in wide distribution throughout the body. The relevant pharmacokinetic parameter is the volume of distribution, which is defined as the theoretical volume of a compartment necessary to account for the total amount of drug in the body if it were present throughout the compartment at the same concentration found in the plasma. High volumes of distribution thus represent extensive tissue binding.

After reaching the systemic circulation, many drugs exhibit a distribution phase during which the drug concentration in each compartment reaches equilibrium, represented by Line B in Figure 39.2. This may be rapid (~15 min for gentamicin) or prolonged (at least six hours for digoxin). There is generally little point in measuring the drug concentration during the distribution phase, especially if the site of action of the drug for clinical or toxic effects lies outside the plasma compartment and the plasma concentration in the distribution phase is not representative of the concentration at the receptor.

Between-patient variations in volume of distribution caused by differences in physical size, the amount of adipose tissue and the presence of disease (e.g. ascites) can significantly weaken the relationship between the dose of drug taken and the plasma concentration in individual patients and hence increase the need for concentration monitoring.

Elimination (metabolism and excretion)

When a drug has been completely distributed throughout its volume of distribution, the pharmacokinetics enter the elimination phase (line C in Fig. 39.2). In this phase, the drug concentration falls owing to metabolism of the drug (usually in the liver) and/or excretion of the drug (usually via the kidneys into the urine or via the liver into the

bile). These processes are described by the pharmacokinetic parameter *clearance*, which is a measure of the ability of the organs of elimination to remove active drug. The clearance of a drug is defined as the theoretical volume of blood that can be completely cleared of drug in unit time (*cf.* creatinine clearance, see Chapter 7). Factors affecting clearance and hence the rate of elimination include body weight, body surface area, renal function, hepatic function, cardiac output, plasma protein binding and the presence of other drugs which affect the enzymes of drug metabolism including alcohol and nicotine (tobacco use). Pharmacogenetic factors may also have a profound influence on clearance.

Rate of elimination is often expressed in terms of the *elimination rate constant* or the *elimination half-life*, which is the time taken for the amount of drug in the body to fall to half its original value and is generally easier to apply in clinical situations. Both the elimination rate constant and the elimination half life can be calculated from the clearance and the volume of distribution (see Pharmacokinetic texts in Further reading).

With some drugs (e.g. phenytoin), the capacity of the clearance mechanism is limited, and once this is saturated there may be large increments in plasma concentrations with relatively small increases in dose. This phenomenon makes these drugs very difficult to use safely without access to concentration monitoring.

Protein binding

Many endogenous constituents of plasma are carried on binding proteins in plasma (e.g. bilirubin, cortisol and other hormones), and drugs are also frequently bound to plasma proteins (usually albumin and α_1 -acid glycoprotein). Acidic drugs (e.g. phenytoin) are, in the main, bound to albumin, whereas basic drugs bind to α_1 -acid glycoprotein and other globulins.

The proportion of drug bound to protein can vary from zero (e.g. lithium salts) to almost 100% (e.g. mycophenolic acid). The clinical effect of a drug and the processes of metabolism and excretion are related to the free concentration of a drug, with drug bound to protein effectively acting as an inert reservoir. Variations in the amount of binding protein present in plasma can thus change the total measured concentration of the drug in the blood without necessarily changing the free (active) concentration, and this is a further reason why measured plasma concentrations may not relate closely to clinical effect. Other drugs or endogenous substances that compete for the same binding sites on protein will also affect the relationship between free and bound drug. For these reasons, it has been suggested that the free concentration of a drug rather than the total (free + protein-bound) concentration should be measured by a TDM service, at least for those drugs exhibiting significant protein binding, e.g. phenytoin. Despite the apparent logic of this idea, it has not been widely adopted because of methodological difficulties and continuing controversy about whether the theoretical benefits are realized in clinical practice. It is nonetheless important to be aware of the effects of changes in protein binding when interpreting total drug concentrations in plasma, especially:

- when the free fraction of a drug may change within an individual over time, e.g. in pregnancy (with falling albumin concentration) or when other drugs bound at the same binding site are added or withdrawn
- when there is a highly abnormal binding protein concentration in plasma; for example, in severe hypoalbuminaemia
- when a pathological state (e.g. uraemia) results in the displacement of drug from the binding sites.

The discussion above has outlined the pharmacokinetic factors that govern the relationship between dose and the drug concentration in the plasma or at the active site, and has indicated how these may vary between patients to produce poor correlation between the dose prescribed and the effective drug concentration at the site of action. In general, once steady-state has been reached, plasma concentrations in the individual should exhibit a constant relationship to the concentration at the site of action governed by the distribution factors discussed above, but this may not be the case if blood supply to target tissues is impaired (e.g. for poorly vascularized tissues or a tumour that has outgrown its blood supply (cytotoxic drugs)) or a site of infection that is not well perfused (antibiotics)).

Pharmacodynamic factors

Pharmacodynamics is the study of the relationship between the concentration of drug at the site of action and the biochemical and physiological effect. The response of the receptor may be affected by the presence of drugs competing for the same receptor, the functional state of the receptor or pathophysiological factors such as hypokalaemia. Interindividual variability in pharmacodynamics may be genetic or reflect the development of tolerance to the drug with continued exposure. High pharmacodynamic variability severely limits the usefulness of monitoring drug concentrations as they are likely to give a poor indication of the effectiveness of therapy.

Which drugs should be measured?

The above discussion provides a basis for determining which drugs are good candidates for therapeutic drug monitoring. As stated at the beginning of this chapter, the aim of TDM is the provision of useful information that may be used to modify treatment. For this reason, it is generally inappropriate to measure drug concentrations where there is a good clinical indicator of drug effect. Examples of this are the measurement of blood pressure during antihypertensive therapy; glucose in patients treated with hypoglycaemic agents; clotting studies in patients treated with heparin or warfarin, and cholesterol in patients treated with cholesterol-lowering drugs. While plasma concentration data for such drugs are valuable during their development to define pharmacokinetic parameters and dosing regimens, TDM is not generally helpful in the routine monitoring of patients. It may have a limited role in detecting poor adherence or poor drug absorption in some cases. However, where there are no such clinical markers of effect or where

symptoms of toxicity may be confused with those of the disease being treated, concentration monitoring may have a vital role.

Therapeutic drug monitoring is useful only for drugs that have a poor correlation between dose and clinical effect (high pharmacokinetic variability). Clearly, if dose alone is a good predictor of pharmacological effect, then measuring the plasma concentration has little to contribute.

However, clinically useful TDM does require that there is a good relationship between plasma concentration and clinical effect. If drug concentration measurements are to be useful in modifying treatment, then they must relate closely to the effect of the drug or its toxicity (or both). This allows definition of an effective therapeutic 'window' – the concentration range between the minimal effective concentration and the concentration at which toxic effects begin to emerge – and allows titration of the dose to achieve concentrations within that window. Demonstration of a clear concentration–effect relationship requires low between-individual pharmacodynamic variability (see above), the absence of active metabolites that contribute to the biological effect but are not measured in the assay system and (usually) a reversible mode of action at the receptor site. Reversible interaction with the receptor is required for the intensity and duration of the response to the drug to be temporally correlated with the drug concentration at the receptor.

Many drugs have active metabolites, and some drugs are actually given as pro-drugs, in which the parent compound has zero or minimal activity and pharmacological activity resides in a product of metabolic transformation. For example, mycophenolate mofetil is metabolized to the active immunosuppressant mycophenolate. It will be clear that useful information cannot be obtained from drug concentration measurements if a substantial proportion of the drug's effect is provided by a metabolite that is not measured and whose concentration relationship to the parent compound is undefined. In some cases (e.g. amitriptyline/nortriptyline), both drug and metabolite concentrations can be measured and the concentrations added to give a combined indication of effect, but this assumes that drug and metabolite are equally active. In other cases (e.g. carbamazepine and carbamazepine 10,11-epoxide), the active metabolite is analysed and reported separately.

Therapeutic drug monitoring is most valuable for drugs which have a narrow therapeutic window. The therapeutic index (therapeutic ratio, toxic-therapeutic ratio) for a drug indicates the margin between the therapeutic dose and the toxic dose – the larger, the better. For most patients (except those who are hypersensitive), penicillin has a very high therapeutic ratio and it is safe to use in much higher doses than those required to treat the patient, with no requirement to check the concentration attained. However, for other drugs (e.g. immunosuppressives, anticoagulants, aminoglycoside antibiotics and cardiac glycosides), the margin between desirable and toxic doses is very small, and some form of monitoring is essential to achieve maximal efficacy with minimal toxicity.

The criteria for TDM to be clinically useful are summarized in [Box 39.1](#).

BOX 39.1 Criteria for clinically useful TDM

- Absence of good clinical markers of drug effect
- Poor correlation between dose and clinical effect (high pharmacokinetic variability)
- Good correlation between plasma drug concentration and clinical effect (low pharmacodynamic variability, no active metabolites)
- Narrow concentration interval between therapeutic and toxic effects

The list of drugs for which TDM is of proven value is relatively small ([Box 39.2](#)). Phenytoin and lithium are perhaps the best and earliest examples of drugs that meet all the above criteria and for which TDM is essential. The aminoglycoside antibiotics, chiefly gentamicin and tobramycin, also qualify on all counts. A number of other drugs that are frequently monitored fail to meet one or more of the criteria completely, and the effectiveness of TDM as an aid to management is therefore reduced. The evidence for the utility of monitoring many drugs is based more on practical experience than well-designed studies. However, for newer agents such as the

BOX 39.2 Drugs for which there is a clear indication for measurement

- Aminoglycoside antibiotics (e.g. gentamicin, tobramycin)
- Carbamazepine
- Ciclosporin
- Digoxin
- Lamotrigine
- Lithium
- Methotrexate
- Phenytoin
- Sirolimus
- Tacrolimus
- Theophylline

Drugs whose measurement may be useful in some circumstances but not generally practised

- Amiodarone
- Antiretroviral drugs (protease inhibitors, NNRTIs)
- β -Blockers
- Caffeine
- Chloramphenicol
- Clozapine
- Disopyramide
- Flecainide
- Flucytosine
- Methadone
- Morphine
- Mycophenolic acid
- Olanzapine
- Phenobarbital
- Procainamide (and N-acetylprocainamide)
- Teicoplanin
- Tricyclic antidepressants (amitriptyline/nortriptyline, imipramine/desipramine, dothiepin)
- Valproate
- Vancomycin

immunosuppressants and antiretroviral drugs, there is good evidence supporting the benefits of TDM in improving clinical outcomes. (Further information is given in Individual Drugs, below.)

USE OF THERAPEUTIC DRUG MONITORING

Once the narrow range of drugs for which TDM can provide useful information has been defined, it should not be assumed that TDM is necessary for every patient on these drugs at every visit. For TDM to be used for maximum patient benefit and optimal cost-effectiveness, six important criteria must be satisfied each time a sample is taken. These are summarized in [Box 39.3](#) and will now be discussed briefly.

Appropriate clinical question

The first essential for making effective use of any laboratory test is to be clear at the start what question is being asked. This is particularly true for TDM requests, and much time, money and effort is wasted on requests where the indication for analysis has not been clearly defined. If the question is uncertain, the answer is likely to be unhelpful.

The two main reasons for monitoring drugs in blood are to ensure effective therapy and to avoid toxicity. Effective therapy requires that sufficient drug reaches the drug receptor to produce the desired response (which may be delayed in onset). Where a drug is prescribed and the desired effect is not achieved, this may or may not be due to insufficient dosage, since there may be other reasons (individual idiosyncrasy, drug interactions etc.) for the lack of effect. Those at the extremes of age – neonates and the very elderly – have metabolic processes that render them differently susceptible even to weight-adjusted doses. For example, in neonates, the metabolism of theophylline is qualitatively as well as quantitatively different from that in older children. In the very elderly, there may be considerable alterations in absorption and also renal clearance. Drug interactions may also produce a reduced clinical effect for a given dose.

In patients on long-term therapy, once a steady-state concentration that produces a satisfactory clinical effect has been obtained, this concentration can be documented as an effective baseline for the individual patient. If circumstances subsequently change, then changes in response can be related back to both the dose and the plasma concentration of the drug. This is particularly important for psychotropic drugs. ‘Baseline’ concentrations

may however change over time as disease processes develop, or gradually with increasing age or changes in drug metabolism.

The avoidance of iatrogenic toxicity is probably the most pressing case for the practice of TDM. The aim is to ensure that drug (or metabolite) concentrations are not so high as to produce symptoms/signs of toxicity. Since a narrow therapeutic index is a prerequisite for drugs suitable for TDM, it is inevitable that toxicity will occur in a small proportion of patients, even with all due care being exercised. Toxicity can never be diagnosed solely from the plasma drug concentration, and it must always be considered in conjunction with the clinical circumstances since some patients will show toxicity when their concentrations are within the generally accepted therapeutic range and others will tolerate concentrations outside the range with few or no ill-effects. The advantage of regular monitoring is that such circumstances may be recorded and the range for that individual adjusted accordingly.

There are two main factors that may lead to inappropriately high plasma drug concentrations. The first is an inappropriate dosing regimen, either due to a single gross error or (more often) a gradual build-up of plasma concentration either because the dose is slightly too high for the individual or because of the development of hepatic or renal insufficiency.

The second factor leading to toxicity is pharmacokinetic drug interactions. Patients are often treated with more than one drug, which can interfere with each other's actions in a number of ways, for example:

- displacement from protein binding sites
- competition for hepatic metabolism
- induction of hepatic metabolizing systems
- competition for renal excretory mechanisms.

Examples of the more commonly encountered drug interactions involving drugs measured for TDM purposes are given in [Table 39.1](#). There are, however, numerous other examples and the analyst faced with an unusual response or an inappropriate plasma concentration should seek a full drug history to assist in determining the explanation. This history should include specific enquiry about the use of alternative therapies, since many herbal and other non-pharmaceutical remedies (which may not be immediately mentioned by patients) may have significant effects on drug concentrations due to the induction of drug metabolizing enzymes. St John's wort (*Hypericum perforatum*), a perennial herb with bright yellow flowers, is one example. It has been shown to induce the hepatic drug-metabolizing enzymes CYP3A4 and CYP2B6, and hence reduce the steady-state concentrations of many drugs.

Therapeutic drug monitoring is particularly useful in confirming toxicity when both under- and overdosage with the drug in question give rise to similar clinical features, for example arrhythmias with digoxin, or fits with phenytoin, and where high drug concentrations give rise to delayed toxicity, for instance high aminoglycoside concentrations, which may give rise to irreversible ototoxicity if prompt action is not taken.

Where a patient is known to be suffering from toxicity and the plasma drug concentration is high, monitoring is often required to follow the fall in concentration

BOX 39.3 Criteria for effective drug monitoring

- Appropriate clinical question
- Accurate patient information
- Appropriate sample
- Accurate analysis
- Relevant clinical interpretation
- Effective action taken

TABLE 39.1 Common drug interactions encountered in therapeutic drug monitoring

Effector	Drug(s) affected (examples)
Interactions due to induction of metabolism	
Phenytoin	Phenytoin, phenobarbital, digoxin
Phenobarbital	Phenobarbital, phenytoin, prednisolone, warfarin, digoxin
Rifampicin	Cortisone, digoxin, methadone, quinidine, theophylline
St John's wort	Ciclosporin, tacrolimus, digoxin, theophylline, antiretrovirals
Interactions due to inhibition of metabolism	
Chloramphenicol	Phenytoin, tolbutamide
Cimetidine	Diazepam, phenytoin, theophylline, warfarin
Ethanol	Diazepam
Erythromycin	Theophylline
Isoniazid	Carbamazepine, phenytoin
Oral contraceptives	Theophylline
Quinidine	Digoxin
Valproate	Phenobarbital
Interactions due to effects on excretion	
Cimetidine, ranitidine	Procainamide
Verapamil	Digoxin
Amiodarone	Digoxin
Spironolactone	Digoxin
Quinidine	Digoxin
Interactions due to displacement from protein binding	
Valproate	Phenytoin
Salicylate	Warfarin, sulphonylureas, methotrexate
Interactions due to enhanced absorption	
Erythromycin	Digoxin

following cessation of treatment. For example, in anti-convulsant overdose, it is important for the clinician to know when drug concentration(s) are likely to reach the therapeutic range, since reinstatement of therapy will then be required to prevent seizures.

Other important reasons for therapeutic drug monitoring include guiding dosage adjustment in clinical situations in which the pharmacokinetics are changing rapidly (e.g. neonates, children or patients in whom hepatic or renal function is changing) and defining pharmacokinetic parameters and concentration–effect relationships for new drugs.

The use of TDM to assess adherence (compliance) is to some extent controversial. Clearly, assessment of adherence could provide justification for monitoring every drug in the pharmacopoeia, at enormous cost and for little clinical benefit. Further, application of TDM in this situation is not simple. A patient with a very low or undetectable drug concentration is usually assumed to be non-compliant, but the situation is much less clear when a concentration that is only slightly low (based on population data) is found. Is the patient non-compliant or are other factors (such as poor absorption, induced metabolism or altered protein binding) leading to the low concentration? Much harm may be done in these situations by assumptions of poor adherence to therapy. Adherence can be assessed in other ways, by tablet-counting, supervised medication in hospital or the use of carefully posed

questions that are non-judgemental, for example ‘How often do you forget to take your tablets?’ Such approaches are likely to be more effective than TDM in detecting and avoiding poor adherence to therapy, though TDM may have a role in patients with poor symptom control, who deny poor adherence despite careful questioning and interventional reinforcement.

Accurate patient information

Proper interpretation of TDM data depends on having some basic information about the patients and their recent drug history. The importance of this requirement has led many laboratories to design specific request forms for TDM. These are now being superseded by computerized request ordering systems that allow collection of essential items of information at the time of making the request, although it remains important to keep the requesting interface user-friendly and not too complex. Intelligent system design is necessary. Basic information requirements for TDM requests are summarized in [Table 39.2](#).

Appropriate sample

The vast majority of TDM applications require a blood sample. In general, serum or plasma can be used, although for a few drugs that are concentrated within red cells, for example, ciclosporin, whole blood is preferable as concentrations are higher and partition effects are avoided. The literature is poor on the differentiation between serum and plasma samples and careful selection of sample tubes is necessary to avoid interferences by anti-coagulants, plasticizers, separation gels etc, either in the assay system or by absorption of sample drug onto the gel or the tube, reducing the amount available for analysis. If in doubt, serum collected into a plain glass tube with no additives is usually safest. Haemolysis should be avoided, as it may cause in vivo interference if the drug is concentrated in erythrocytes, or in vitro interference in immunoassay systems. For drugs that are minimally protein bound, there need be no restrictions on the use

TABLE 39.2 Information requirements for TDM requests

	Essential	Desirable
Patient	Name Age	Weight Renal/hepatic function
	Gender Hospital/health system ID number	
Problem	Pathology/clinical details Reason for request (e.g. poor response, ?toxic)	
Therapy	Drug of interest dose formulation and route of administration duration of therapy date/time last given	Other drugs – list all

of tourniquets; however, stasis should ideally be avoided for drugs which are highly bound to albumin, e.g. phenytoin. Care should also be taken to avoid contamination with local anaesthetics, for example lidocaine, which are sometimes used before venepuncture. Where intravenous therapy is being given, care must be taken to avoid sampling from the limb into which the drug is being infused.

Urine is of no value for quantitative TDM. Saliva may provide a useful alternative to avoid venepuncture (especially in children) or when an estimate of the concentration of free (non-protein-bound) drug is required. Saliva is effectively an *in vivo* ultrafiltrate and concentrations reflect plasma free drug concentrations quite well for drugs that are essentially unionized at physiological pH. Salivary monitoring is unsuitable for drugs that are actively secreted into saliva (e.g. lithium) and drugs that are strongly ionized at physiological pH (e.g. valproic acid, quinidine) as the relationship between salivary and plasma concentrations becomes unpredictable. Careful collection of the sample is required, and the mouth should be thoroughly rinsed with water prior to sampling. Sapid (taste related) or masticatory (chewing on an elastic band) stimulation of saliva flow is used to increase volume. The mucoproteins in saliva make the sample difficult to handle, and centrifugation is usually necessary to remove cellular debris. These and other problems have meant that salivary analysis has not been widely adopted for routine TDM.

It is normally necessary for the patient to be at steady-state on the current dose of drug, i.e. when absorption and elimination are in balance and the plasma concentration is stable. This is true except when suspected toxicity is being investigated, when it is clearly inappropriate to delay sampling until steady-state has been reached. The time taken to attain the steady-state concentration is determined by the plasma half-life of the drug, and the relationship between the number of half-lives which have elapsed since the start of treatment and the progress towards steady-state concentrations is shown in Table 39.3.

It is frequently stated that five half-lives must elapse before plateau concentrations are achieved, unless loading doses are employed when they are achieved much more rapidly. As shown in Table 39.3, the plasma concentration after 3.3 half-lives is 90% of the predicted steady-state concentration, and this may be taken as the minimum time for sampling for routine purposes after starting a drug or changing the dose. For drugs with a long half-life (e.g. digoxin or phenobarbitone), two weeks

or more may be required before steady-state is achieved, especially if the drug is renally excreted and renal function is poor.

In neonates, the rapidly changing clinical state, degree of hydration and dosage requirements make the concept of steady-state a theoretical ideal rather than an attainable goal and there is little value in delaying measurements in the hope of attaining steady-state.

A further requirement for many drugs is for samples to be taken at the appropriate time following the last dose. The size of the fluctuations in plasma concentration between doses depends on the dosage interval and the half-life of the drug. Frequent dosing avoids large peaks and transient toxic effects but is unpopular with patients, difficult to comply with and more likely to lead to medication errors. Less frequent dosing can give rise to large fluctuations in plasma concentration. To some extent, these opposing considerations can be reconciled with the use of sustained-release preparations.

There is no single optimum time for taking samples in relation to dose. The most reproducible for sampling is immediately (<30 min) before a dose (trough concentration), when the lowest concentrations in the cycle will be obtained. This is the optimum time if an indication of drug efficacy is required, and will show least between-sample variation in patients on chronic therapy. The use of peak and trough concentrations for detecting toxicity of aminoglycoside antibiotics has become less relevant with the advent of once-daily dosing regimens, but sampling at two hours post-dose for ciclosporin can give an excellent estimate of the probable area under the curve and has become a common and effective TDM technique, though sampling in the middle of the absorption/distribution phase in this way does require very accurate standardization of sampling time in relation to the last dose. In the case of digoxin, a specimen should not be taken within 6 h of the dose, since the digoxin absorption and distribution peak may be extremely sharp and high (Fig. 39.3) and serum digoxin concentrations in this period do not reflect tissue concentrations.

TABLE 39.3 Percentage of steady-state plasma concentration attained at various times after starting/changing therapy

Time (drug half-lives)	% steady-state
0.5	29
1.0	50
2.0	75
3.0	88
3.3	90
4.0	94
5.0	97
7.0	99

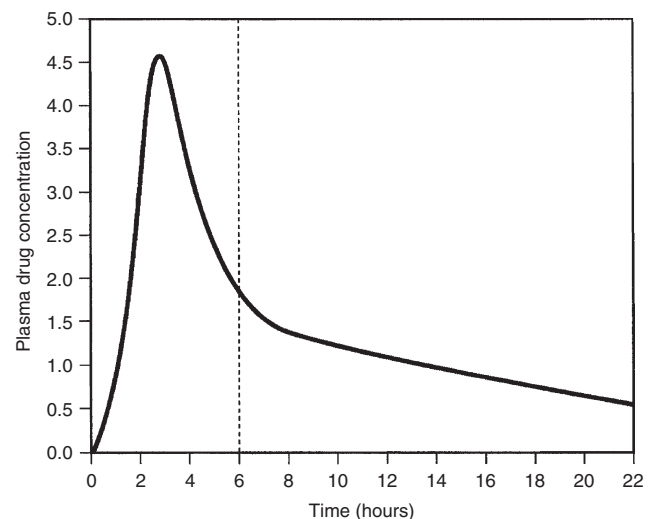


FIGURE 39.3 ■ The concentration/time curve for a patient given digoxin orally. Note the sharp peak and relatively stable plateau concentration after 6 h.

Further details on optimal sampling times for individual drugs are given in the relevant sections of this chapter. When computer programmes are used for pharmacokinetic parameter optimization and dosage prediction, it becomes extremely important that dosage and sampling times are accurately known, as large prediction errors can result from inaccurate data.

Accurate analysis

Establishing a relationship between plasma drug concentrations and effect requires accurate, precise and reliable analytical methods. The choice of methods is vast, encompassing immunoassays (isotopic and optical), chromatography (e.g. gas liquid chromatography (GLC) and high performance liquid chromatography (HPLC)), increasingly with mass spectrophotometric detectors) and a variety of novel techniques involving dry chemistry.

Selecting the most appropriate analytical methods is often challenging, and the choice depends on the availability of staff, expertise and equipment, the anticipated batch sizes, the range of drugs to be measured and the turnaround time required. There is no 'best fit' for every situation. Some guidelines are given in Table 39.4. Assays used in routine TDM must be specific, showing no interferences from endogenous compounds, drug metabolites or exogenous sources, and must also work with small sample sizes (certainly <1 mL plasma/serum and ideally 10–100 μ L).

Most analytical methods determine the total concentration of drug in plasma or serum. As discussed above, it is sometimes useful to obtain an estimate of the free (non-protein-bound) drug concentration. The main methods for separation of bound and free drug prior to quantitation by conventional assays are equilibrium dialysis, ultrafiltration and centrifugation through a fine-mesh membrane. The latter two are non-equilibrium techniques and are therefore much quicker than equilibrium dialysis, which generally requires an overnight incubation. It is often difficult to assess the reliability of different procedures as no single method gives consistently accurate estimates of free drug concentration.

TABLE 39.4 Guidelines for choice of methods for therapeutic drug monitoring analyses

Purpose for which required	Most appropriate method(s)
On-site analyses in clinics	Point of care immunoassay devices
Urgent analyses	Optical immunoassays
Batch analyses, single drug	Optical immunoassays, HPLC, GC, HPLC/mass spectrometry (MS)
Batch analyses, multiple drugs	HPLC, GC (optical immunoassays become uneconomic)
Single analyses with metabolite patterns	HPLC/MS or LC/MS

For some drugs, no choice of method exists.

Relevant clinical interpretation

In the 40 years or so that TDM has been practised routinely, it has been shown that having drug concentration measurements available to clinicians does not in itself result in improved clinical care. Improved outcome depends on application of the analytical result to a specific clinical situation with appropriate expertise.

In particular, it is important to understand that the widely quoted (and just as widely misused) 'therapeutic ranges' for drugs represent a guide to the approximate concentrations that produce a therapeutic response in the majority of patients, rather than a set of inflexible concentration limits to which dosage regimens must be directed. 'Target ranges' has been suggested as a better term, which at least implies that these are aims, rather than that all concentrations within the specified range are therapeutic (and all outside are not). Many patients need plasma drug concentrations above the upper limit of the target range for effective therapy, and such concentrations must not provoke knee-jerk dosage reduction. Specialist clinicians usually appreciate this fact, but non-specialists frequently do not, and laboratory staff or pharmacists have an essential educational role here. Conversely, plasma drug concentrations below the lower limit of the target range may produce a satisfactory response in some patients, and arbitrary dose increases will merely increase the likelihood of toxicity without added benefit. In one of the earliest papers on TDM, Koch–Weser wrote: 'Therapeutic decisions should never be based solely on the drug concentration in the serum'. The cardinal principle, often repeated but still forgotten, is to treat the patient, not the drug concentration.

Effective action taken

As implied above, drug concentrations above the target range do not invariably require a reduction in dosage. If the patient is symptom-free, a careful search for signs of toxicity should be made. If there is no evidence of toxicity, the patient may be best served by doing nothing, although for some drugs (e.g. phenytoin) continued monitoring for the development of long-term undesirable effects is advisable. Similarly, drug concentrations below the target range in a patient who is well and free from symptoms do not necessarily require an increased dose, although in some cases (e.g. digoxin), they may provide evidence that the drug is no longer effective, and that stopping it under medical supervision may be worth trying.

For most drugs, there is a relatively linear relationship between dose and plasma concentration at steady-state, the major exception to this being phenytoin (see later); thus doubling the dose should double the plasma concentration. The difficulty, therefore, is in determining the initial dose to use, especially where a rapid effect, for example the prevention of cardiac arrhythmias, is required. The aim then is to achieve a steady-state concentration rapidly, without waiting for five half-lives to elapse, particularly for drugs with long half-lives. For this reason a 'loading' dose of several times the daily maintenance dose is given, effectively reaching steady-state over two half-lives (Fig. 39.4). The calculation of the doses required for

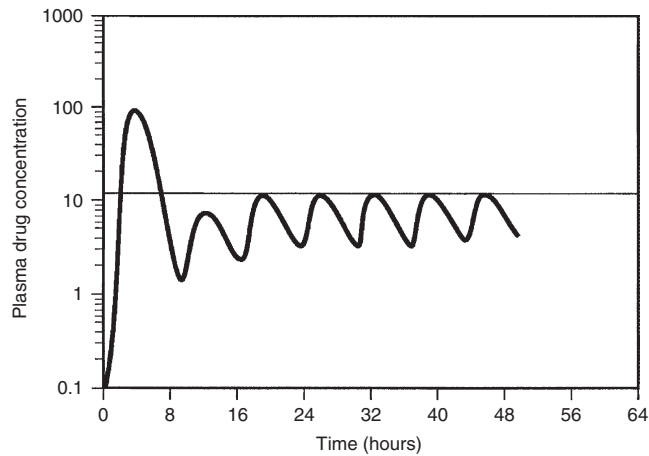


FIGURE 39.4 ■ The effect of a loading dose on the rate of achieving a steady-state concentration

loading and maintenance of a steady-state concentration is relatively straightforward and the relevant equations are given in [Appendix 39.1](#).

In practice, dosage prediction can be approached in two main ways. It can be estimated from average population parameter values using the equations in [Appendix 39.1](#), or the measurement of one or more drug concentrations at appropriate times in the specific individual allows calculation of patient-specific parameters, which then allows further dosage optimization – TDM-assisted dosage optimization. The details are beyond the scope of this chapter and specialist pharmacokinetic texts (see [Further reading](#)) should be consulted. Numerous nomograms and software packages are available for dosage optimization based on TDM data and Bayesian principles. It is vital to ensure that the baseline data are accurate, for example ensuring that the time the drug was actually taken in relation to the sampling time is known (rather than when it was supposed to be taken).

PROVISION OF A THERAPEUTIC DRUG MONITORING SERVICE

The basic essentials for an effective TDM service are the availability of appropriate analytical methods and quality assurance procedures, specialist expertise and the ability to produce results within a clinically relevant time.

Staff

A TDM service requires experienced analysts who understand the basis of the procedures involved and are competent to advise on the analytical sensitivity and specificity of the methods being used and to maintain exacting standards of accuracy, precision and document control. In addition, it is essential that the service has access to specialist clinical advice from someone who understands the principles of pharmacokinetics and therapeutics – an appropriately trained laboratory scientist, a pharmacist or a specialist clinician. The best service is generally obtained with a team approach in

which pharmacists, clinical scientists, technical and medical staff all play a role. Difficult cases can then be discussed before any samples are taken and advice obtained on whether concentration monitoring is indicated and on the appropriate timing of any samples required. Following analysis, results are scrutinized by members of the team and appropriate recommendations for action conveyed to the clinician.

However, this ideal degree of involvement is impractical in many hospital situations, and even more difficult for outpatient or primary care work. Provided the request form contains the necessary clinical information (see above), laboratory staff can check these details on receipt of the specimen and advise if the specimen or the reason for analysis is appropriate. Once the analysis is complete, all results (or just those which fall outside defined limits) can be assessed by specialist staff and communicated to the requesting clinician with the appropriate degree of urgency and with recommendations for action where appropriate.

Turnaround time

The ideal turnaround for TDM analyses requires that results are returned to the clinician before the next dose of the drug is due, in order that adjustments can be made immediately. This is not feasible in cases of three or four times daily dosage regimens but, provided that a responsive service is available, patients should not be exposed to inappropriate dosage regimens for more than 24 h after the specimen was taken.

Point-of-care testing

Point-of-care testing is appropriate in some clinical situations since the availability of a plasma drug concentration at the time of consultation allows the clinician to make immediate decisions on dosage adjustments based on objective data. The savings in patient time and improvements in clinic throughput have been shown to outweigh the extra costs of on-site testing in a number of settings. Several methods are available for on-site TDM testing, but need to be carefully evaluated to ensure that their analytical performance is adequate in the setting in which they are used.

Reporting

The reporting of a plasma concentration alone is of limited value to the clinician who is not fully familiar with TDM. Reports that contain information on the reason for the request, the dosage and timing along with the result and appropriate interpretative comments are much more useful. All urgent results should be telephoned and the fact noted. The provision of cumulative reports improves liaison between the laboratory and the clinician, and has been shown to improve patient care.

Units

There has been considerable debate about the appropriate units (mass or molar) for reporting measurements of

drug concentration. Most of the toxicological literature has been in mass units, but molar units have theoretical advantages in making it easier to equate drug and metabolite concentrations. Uniformity of reporting is important for patient safety reasons, as confusion over units can result in dangerous misinterpretations. Within the UK, at least, there is now agreed consensus that SI mass units with the litre as the unit of volume should be used for all therapeutic drug assays, except for a specified few which have always been reported in molar units (methotrexate, lithium, thyroxine and iron).

Quality assurance

Measurements of plasma drug concentrations can only be used effectively to diagnose toxicity or under-dosage and to monitor treatment if the measurements are accurate and reproducible. The need for effective internal quality control and external quality assessment procedures for all methods used to produce results is therefore as great as in other areas of clinical biochemistry. Analytical laboratories should be accredited to ISO 15189.

Continuing education

A good TDM service provides an environment for education of trainee staff, not only from biochemistry but from clinical pharmacy and clinical pharmacology. The involvement of staff from other disciplines and their exposure to the analytical and interpretational aspects should be encouraged. Insistence on the provision of full clinical information and the feedback of interpretation on correctly documented cases can lead to considerable improvements in the effectiveness of the service.

PHARMACODYNAMIC MONITORING, BIOMARKERS AND PHARMACOGENETICS

Classic TDM uses drug concentration measurements in body fluids to guide optimization of therapy and minimize adverse consequences. In recent years, other methods of guiding drug therapy have been introduced, and though they do not fit the strict definition of TDM they merit discussion as they are becoming increasingly relevant to the provision of effective drug therapy.

Pharmacodynamic monitoring is the study of the biological effect of a drug at its target site, and has been applied to areas of immunosuppressive therapy and cancer chemotherapy. For example, the biological effect of the immunosuppressive calcineurin inhibitors ciclosporin A and tacrolimus can be assessed by direct measurement of calcineurin phosphatase activity. The main disadvantage of pharmacodynamic monitoring that has emerged to date is that the assays involved are often significantly more complex and time-consuming than the measurement of a drug by chromatography or immunoassay.

Any biochemical measurement that can be used to determine efficacy, extent of toxicity or individual pharmacodynamics for a therapeutic agent is called a *therapeutic biomarker*. Hitherto, most such markers have been markers of toxicity rather than of therapeutic efficacy

(e.g. urinary N-acetylglucosaminidase as an index of tubular damage caused by nephrotoxic drugs), but there is increasing interest in biomarkers that give direct information about drug efficacy, for example red cell 6-thioguanine nucleotide concentrations in the assessment of thiopurine drug efficacy or measurement of inosine 5'-monophosphate dehydrogenase (IMPDH) activity as a marker of the IMPDH inhibitor mycophenolic acid. Biomarker monitoring can provide an integrated measure of all biologically active species (parent drug and metabolites) present, so that therapeutic ranges can be defined more closely. It is also often free from the matrix and drug disposition effects that frequently complicate drug concentration measurements.

Pharmacogenetic studies (studies of genetic influences on pharmacological responses) have wide-ranging clinical relevance. The enzymes responsible for metabolizing drugs exhibit wide interindividual variation in their protein expression or catalytic activity, resulting in quantitative and qualitative differences in drug metabolism between individuals. This variation may arise from transient effects on the enzyme, such as inhibition or induction by other drugs or metabolites, or may result from specific mutations or deletions at the gene level. Pharmacogenetic polymorphism is defined as the existence in a population of two or more alleles at the same locus that result in more than one phenotype with respect to the effect of a drug, the rarest of which occurs with a significant frequency (usually taken as 1%). The term *pharmacogenomics* is used to describe the range of genetic influences on drug metabolism, and the application of this information to the practice of tailoring drugs and dosages to specific individuals to enhance safety and/or efficacy. This practice, often referred to as 'Personalized Medicine', is a major growth area for 21st century medicine.

Determination of an individual's ability to metabolize a specific drug may be performed either by administering a test dose of the drug or a related compound and measuring the metabolites formed (phenotyping), or by specific genetic analysis (genotyping). The information obtained can dramatically improve the clinician's ability to select a drug dose appropriate to the specific requirements of the individual. For example, many of the isoenzymes of the cytochrome P450 superfamily responsible for drug oxidation show genetic polymorphisms that affect the extent of drug metabolism and produce differences in clinical response. The CYP2D6 isoform has more than 100 allelic variants, and metabolizes about a quarter of all drugs used in medicine, including many antiarrhythmics and psychoactive drugs. The rate of metabolism of a test dose of debrisoquine or dextromethorphan has been widely used for the determination of CYP2D6 phenotype and the differentiation of poor metabolizer (PM), extensive metabolizer (EM) and ultra extensive metabolizer (UEM) phenotypes. Alternatively, genetic analysis can be used to define the CYP2D6 genotype and identify the alleles associated with the PM phenotype (of which the most common are CYP2D6*3, *4, *5 and *6). Once determined, the phenotype can be used to guide dosing for any of the wide range of drugs metabolized by the CYP2D6 isoform, and ensure that lower doses are used for individuals with the PM phenotype, avoiding toxicity. The PM phenotype

is found in 7–10% of Caucasians but less frequently in people of Asian origin.

Pharmacogenetic data have many other clinical applications, including anticoagulation (polymorphisms of the CYP2C9 isoform and the *VKORC1* genes in assessing susceptibility to warfarin), oncology/immunosuppression (thiopurine methyltransferase polymorphisms and azathioprine therapy), psychiatry (CYP2D6 isoforms and rate of metabolism of some antidepressants), epilepsy, pain control and other areas. (See Chapter 43 for a further discussion of this topic.)

Integrating information

Pharmacogenetics and new biomarkers provide invaluable adjuncts to conventional drug concentration monitoring, and promise to deliver the ability to create individualized therapeutic regimens. However, integrating the information from all three strands is complex, and will require developments in decision support software and effective strategies for presenting the information in an accessible and comprehensible format to those involved in the provision of care. Pre-treatment pharmacogenetic profiling will allow identification of individuals who are likely to be particularly susceptible or resistant to a proposed regimen, allowing better selection of drug and initial dose. However, the effect of factors such as disease, age or drug interactions means that pharmacogenetics can never tell the whole story, and measuring the concentrations of drugs, metabolites or other biomarkers will still be necessary to complete the picture and deliver truly personalized medicine.

INDIVIDUAL DRUGS

Analgesic/anti-inflammatory drugs

Aspirin (acetylsalicylic acid)

Aspirin (acetylsalicylic acid) is a widely available drug used as a self-prescribed analgesic and as a prophylactic (at low dose) for thromboembolic disease. It is prescribed in relatively high doses for inflammatory arthropathies, since it has both analgesic and anti-inflammatory properties. Aspirin causes gastric erosions and, for this reason, is available in a variety of formulations designed to reduce this side-effect. Once absorbed, aspirin is rapidly converted to salicylate. Salicylate metabolism is complex and variable and its excretion is highly dependent upon urinary acidity, hence the concurrent ingestion of antacids can lead to lower plasma concentrations. The commonest symptom of toxicity is tinnitus, which occurs in the majority of patients once the plasma concentration exceeds 400 mg/L (2.9 mmol/L). Plasma salicylate concentrations may need to be monitored in a minority of patients to avoid toxicity. When used as an anti-inflammatory agent, plasma concentrations should lie between 150 and 300 mg/L (1.1–2.2 mmol/L), measured just before a dose. The half-life of salicylate varies widely between 3 and 20 h, depending on the dose and duration of therapy.

Value of monitoring: low.

Antiarrhythmics and cardiac glycosides

Amiodarone

Amiodarone is used in the treatment of intractable arrhythmias. It can be given intravenously or orally. It is strongly tissue bound and has a very high volume of distribution (about 5000 L), leading to a plasma half-life of approximately 50 days. There is considerable interpatient variability. Effective combined plasma concentrations of amiodarone and its active metabolite, desethylamiodarone, are in the range 0.5–2.5 mg/L (0.7–3.7 µmol/L). Most patients do not require concentration monitoring, but in some patients it may be useful to differentiate treatment failure from poor adherence or suboptimal dosing, while in others it may assist in the confirmation of concentration-related side-effects.

Value of monitoring: moderate.

Digoxin and digitoxin

The cardiac glycosides digoxin and digitoxin increase the force of cardiac contraction and increase cardiac output, and are used in the treatment of cardiac failure. They are also useful in the management of certain supraventricular arrhythmias. Digitoxin is structurally and functionally similar to digoxin, but is less frequently used in the UK.

Digoxin has a long half-life (20–60 h) and is usually given once daily. Clearance is primarily by the kidneys and a reduced glomerular filtration rate (GFR) can lead to toxicity. Effective plasma concentrations are 0.5–2.0 µg/L (0.6–2.6 nmol/L), with concentrations at the lower end of this range (0.5–1.0 µg/L; 0.6–1.3 nmol/L) recommended for the treatment of heart failure. Blood samples should be taken at least 6 h post-dose to allow time for distribution into tissues. Digoxin acts by inhibition of the Na⁺,K⁺-ATPase pump in the membranes of cardiac myocytes. Studies on red blood cells have shown that patients on long-term therapy adapt by production of additional pump activity. As a result, the interpretation of plasma concentrations is different depending upon whether therapy is newly instituted or long term. There is no clear relationship between digoxin concentration and therapeutic effect, but toxicity is a major problem and is more likely at plasma concentrations >2.3 µg/L (3.0 nmol/L). Toxicity at any digoxin concentration is inversely proportional to plasma potassium concentration and, in the presence of hypokalaemia, toxicity may be observed at plasma digoxin concentration as low as 1.2 µg/L (1.5 nmol/L). The patient's age and the severity of heart disease are independent risk factors for development of toxicity. Although the most common presentation of digoxin toxicity is bradycardia, there may also be other cardiac arrhythmias which, coupled with underlying cardiac pathology, may give rise to a confused clinical picture that may be clarified by the measurement of digoxin concentrations. Significant numbers of patients die with digoxin concentrations >2.3 µg/L. The possible contribution of digoxin to the cause of death is not always suspected. It should be noted that, if a patient is treated with anti-digoxin antibodies to reverse toxicity (see Chapter 40), further measurements of

digoxin concentration will not be possible using the commonly available immunoassays until the antibodies have been cleared, which may take up to a week.

It should be noted that in some patients, notably neonates, there exist in the plasma digoxin-like interfering substances (DLIS) that can lead to falsely elevated 'digoxin' concentrations when measured by immunoassay. Elevated DLIS concentrations are encountered in patients with volume-expanded conditions such as uraemia, essential hypertension, liver disease and pre-eclampsia, suggesting that DLIS may be a hormone involved in natriuresis. Falsely low digoxin values due to the presence of DLIS have also been reported. Commercially available immunoassays for digoxin vary in their specificity for DLIS and their ability to differentiate DLIS from exogenous digoxin. The presence of DLIS should be considered if digoxin concentrations are unexpectedly high.

Value of monitoring: moderate to high.

Disopyramide

Disopyramide is effective in controlling supraventricular and ventricular arrhythmias after myocardial infarction. Its use is limited by anticholinergic side-effects including dry mouth, blurred vision and difficulty in urination, and it is contra-indicated in prostatic enlargement. Monitoring can be helpful in ensuring efficacy and avoiding toxicity, though variable protein binding complicates the picture. Monitoring of free disopyramide concentration is recommended, although the assay is not always easily available. Target ranges for trough values are 1.5–5.0 mg/L (4.4–14.7 $\mu\text{mol/L}$) for total disopyramide and 0.5–2.0 mg/L (1.5–5.9 $\mu\text{mol/L}$) for free disopyramide.

Value of monitoring: moderate.

Flecainide

Flecainide is used to control serious symptomatic arrhythmias, junctional re-entry tachycardia and paroxysmal atrial fibrillation. It is eliminated via the kidneys (40–45%) and metabolized in the liver, where it is a substrate for the CYP2D6 isoform. The half-life is shortened and plasma concentrations are lower in extensive metabolizers. The usual target range is 0.2–1.0 mg/L (0.5–2.4 $\mu\text{mol/L}$), with most patients responding in the lower half of the range.

Value of monitoring: moderate.

Procainamide

Procainamide is used to control ventricular arrhythmias after myocardial infarction and atrial tachycardia. It is metabolized in the liver to N-acetylprocainamide (NAPA), which is also active, though slightly less potent. The rate of acetylation of procainamide to NAPA is genetically determined with a bimodal distribution, patients being classified as either slow or fast acetylators. Procainamide has a relatively short half-life (2–5.5 h) and is rapidly cleared. N-acetylprocainamide has a longer half-life (6–12 h) than the

parent drug and renal impairment can result in the accumulation of toxic concentrations of the metabolite. Monitoring of procainamide should therefore include quantitation of both procainamide and NAPA. Trough samples are preferable. Procainamide concentrations should fall between 4 and 10 mg/L (17–42 $\mu\text{mol/L}$). Procainamide and NAPA concentrations combined should not exceed 30 mg/L (100 $\mu\text{mol/L}$).

Value of monitoring: moderate.

Anticonvulsants (antiepileptics)

Carbamazepine/oxcarbazepine

Carbamazepine is a drug of choice for the treatment of simple and complex focal seizures and generalized tonic-clonic seizures. It is also used for the prophylaxis of bipolar affective disorder and mania, and for pain relief in trigeminal neuralgia. Metabolism is primarily by hepatic oxidation. The major metabolite, the 10,11-epoxide, is active but is seldom measured routinely, which complicates the interpretation of plasma carbamazepine concentrations. The best available information suggests that carbamazepine, when prescribed alone, is effective at plasma concentrations of 4–12 mg/L (17–50 $\mu\text{mol/L}$). Nystagmus, ataxia and drowsiness may occur at concentrations >12 mg/L and more serious toxicity >15 mg/L (60 $\mu\text{mol/L}$). Protein binding is variable and drug concentrations in saliva have been studied but have not proved to be widely useful. A troublesome erythematous rash occurs in about 4% of patients, and other chronic side-effects including hepatic and haematological problems may occur, although these are not dose related. Carbamazepine induces its own metabolism, and therapy needs to be initiated at a low dose and built up slowly over weeks. The unpredictable relationship between dose and concentration, the narrow therapeutic index and the presence of numerous drug interactions that alter the pharmacokinetics support the requirement for therapeutic drug monitoring until stable therapy is established.

Oxcarbazepine is also licensed for the same indications as carbamazepine. It is effectively a pro-drug for 10-hydroxycarbazepine, which is active. The role of therapeutic monitoring has not been established.

Value of monitoring: high (carbamazepine); low (oxcarbazepine).

Ethosuximide

Ethosuximide is used in the treatment of absence seizures and is administered once daily as it has a long half-life. In most patients, therapy can be optimized on the basis of clinical response and EEG checks. Toxicity is readily recognizable clinically as anorexia, nausea, vomiting and dizziness. There is therefore no real need to measure plasma concentrations except in patients receiving multiple therapy to decide which drug is responsible for symptoms of toxicity. The therapeutic range is usually taken as 40–100 mg/L (280–710 $\mu\text{mol/L}$) with toxicity likely at concentrations >150 mg/L (1060 $\mu\text{mol/L}$).

Value of monitoring: low to moderate.

Phenobarbital/primidone

Phenobarbital (phenobarbitone) is an anticonvulsant drug in its own right and also the active metabolite of primidone. It is effective but may be sedative in adults and cause behavioural disturbances in children, and is therefore not a recommended first- or second-line drug. Phenobarbital has a relatively long half-life and is cleared by both the renal and hepatic routes. There is a wide inter-individual variation in handling and, for this reason, there are no well-defined therapeutic limits for plasma phenobarbital concentrations. The majority of well-controlled patients have plasma concentrations of 10–40 mg/L (40–160 $\mu\text{mol/L}$), although some individuals may require much higher concentrations owing to the development of tolerance. The toxic symptoms associated with inappropriately high dosage progress from increasing drowsiness to coma. Primidone therapy can be monitored by the phenobarbital concentration if required.

Value of monitoring: moderate (phenobarbital); low (primidone).

Phenytoin

Phenytoin is still widely used and is effective for the treatment of tonic-clonic and focal seizures, but is no longer a first- or second-line drug for the treatment of epilepsy. It is well absorbed, has a long elimination half-life and is cleared mainly by oxidation in the liver. Its enzymatic removal is, however, saturable and the concentration at which this occurs varies between individuals. Once saturation has occurred, there is an exponential (zero order) increase in plasma concentration (Fig. 39.5); saturation can occur within the therapeutic range, so small increments in dose can lead to large increments in plasma

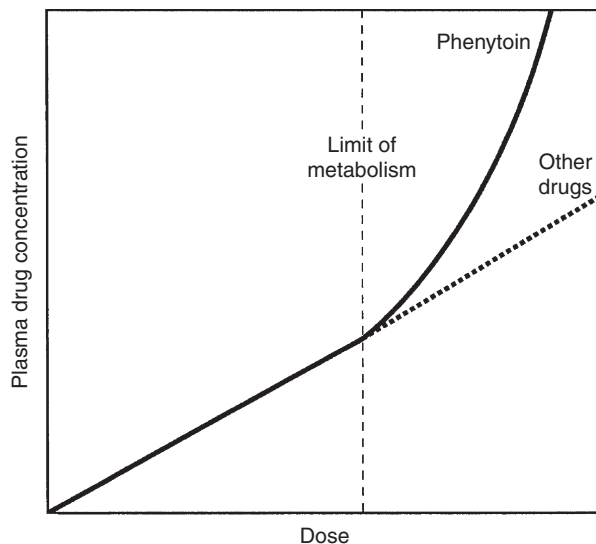


FIGURE 39.5 ■ The relationship between dose and plasma concentration for phenytoin as compared with most other drugs (dotted line). The point where hepatic metabolism is saturated (at which divergence occurs) is specific to the individual.

concentration and severe toxicity. In addition, phenytoin induces its own metabolism, leading to increased clearance and lower plasma concentrations after some weeks. Interaction with other drugs, particularly valproate, can also cause changes in clearance. The drug is highly protein-bound, and, at a given plasma total drug concentration, the lower the plasma protein concentration, the higher will be the free (effective) drug concentration. The unpredictable relationship between dose and plasma concentration, the narrow therapeutic index and the many clinically significant drug interactions mean that monitoring is essential for safe and effective use of the drug. The usual target range is 5–20 mg/L (20–80 $\mu\text{mol/L}$) but higher concentrations may be required in severe epilepsy. Neurotoxicity (nystagmus, dysarthria, ataxia) is usually concentration-dependent, though signs of chronic toxicity, for example gum hyperplasia, are not so clearly linked to concentration.

Value of monitoring: high.

Valproate

Sodium valproate is a popular drug because it does not give rise to drowsiness or the CNS side-effects associated with other anticonvulsants, although it carries a higher risk of congenital malformation and should be used with caution in women of childbearing age. It is recommended as a first-line drug in most types of epilepsy and is used widely in children. It is also used in the treatment of bipolar disorder. However, there is no well-documented relationship between plasma valproate concentration and relief from seizures; indeed, seizures appear to be absent for some time following a dose of valproate when plasma concentrations are too low to measure. A number of cases of severe hepatocellular damage, some fatal, have been reported in patients receiving valproate alone. These appear to be idiosyncratic and are not related to plasma concentration. There is, therefore, no reason to measure plasma valproate in order to avoid hepatotoxicity; however, where the risk of liver damage is high, for example in young children or those with seizure disorders associated with metabolic or degenerative disorders, learning disorders or organic brain disease, baseline and follow-up liver enzyme measurements are indicated.

Valproate is highly protein bound (90–95%) and, when given together with other anticonvulsant drugs, for example phenytoin, can give rise to a (transient) increase in the concentration of free phenytoin. There may, on occasion, be some value in analysing valproate in order to adjust phenytoin or carbamazepine therapy or in the management of complex regimens, but the vast majority of valproate measurements are an unnecessary drain on resources.

Value of monitoring: low.

Newer anticonvulsant drugs

There is now a plethora of newer antiepileptic drugs such as clobazam, clonazepam, felbamate, gabapentin, lamotrigine, levetiracetam, piracetam, retigabine,

rufinamide, topiramate, vigabatrin and zonisamide. Data on concentration–effect relationships are limited, and TDM is not widely used outside specialist centres, since there are no accepted target ranges and most have a wide therapeutic range with overlap of toxic and non-effective concentrations. Tentative target ranges are available in the literature (see [Further reading](#)). Routine monitoring cannot be recommended at present, except for monitoring of lamotrigine concentrations on combination therapy because of the marked effect of drug interactions.

Value of monitoring: low to moderate.

Antidepressants and antipsychotic drugs

Tricyclic antidepressants (amitriptyline, clomipramine, dosulepin, doxepin, imipramine, lofepramine, nortriptyline, trimipramine)

The tricyclic antidepressants show wide interindividual genetic differences in metabolism, and for many of them a good relationship between plasma concentration and clinical effectiveness has been shown. Their use is diminishing in favour of selective serotonin release inhibitors (SSRIs), which are better tolerated and less toxic in overdose. Monitoring is potentially valuable to reduce the incidence of toxicity in susceptible patients.

Value of monitoring: moderate.

Selective serotonin release inhibitors (SSRIs) (citalopram, escitalopram, fluoxetine, fluvoxamine, paroxetine and sertraline)

Therapeutic drug monitoring for SSRIs is of little clinical importance in routine practice, primarily because of their much lower toxicity.

Value of monitoring: low.

Lithium

The lithium cation is of great clinical use in the management of bipolar disorder. Lithium is, however, extremely toxic, with a low therapeutic index. It is prescribed orally as uncoated lithium carbonate or as a slow-release preparation. The usual therapeutic range is 0.4–1.0 mmol/L in samples taken 12 h post-dose, with an upper limit of 0.8 mmol/L in the elderly. Concentrations up to 1.2 mmol/L may be required for the acute treatment of mania in younger patients. Since lithium is nephrotoxic and is excreted renally, excessive dosing can produce a vicious circle in which plasma concentrations increase, causing renal damage, excretion is reduced and concentrations increase further. Plasma concentrations >1.5 mmol/L require intervention, and concentrations >3.0 mmol/L on chronic therapy are potentially fatal. High concentrations in acute overdose are less serious than on chronic therapy, since the tissues are not saturated and distribution will lead to a rapid fall in plasma concentration.

Patients newly started on lithium should be monitored regularly until the relationship between dose and plasma concentration is established. Thereafter, monitoring can be less frequent. Lithium competes with sodium for reabsorption in renal tubules, and alterations in sodium balance or fluid intake may precipitate toxicity. Patients on lithium who start diuretics, or develop diarrhoea and vomiting or renal problems should be assessed and their serum lithium measured. Lithium should not be stopped completely since acute withdrawal can give rise to severe psychiatric symptoms. Thyroid dysfunction (most often hypothyroidism), hyperparathyroidism and nephrogenic diabetes insipidus are recognized side-effects of lithium therapy. Lithium has a half-life of 20–40 h depending on the duration of treatment; steady-state concentrations are therefore obtained approximately seven days after the start of therapy.

Value of monitoring: high.

Other antidepressants

There is no good evidence for a significant relationship between drug concentration and therapeutic outcome for the tetracyclic antidepressants maprotiline, mianserin and mirtazapine, the monoamine oxidase inhibitors moclobemide and tranlycypromine or for trazodone and reboxetine.

Value of monitoring: low.

Antipsychotic drugs

The first-generation antipsychotic drugs such as haloperidol and the phenothiazines chlorpromazine, fluphenazine and perphenazine show marked variation in metabolism between individuals according to their CYP2D6 genotype. Drug and metabolites accumulate in poor metabolizers and overdosing leading to extra-pyramidal side effects and irreversible tardive dyskinesia may occur. If CYP2D6 genotyping is not available, TDM may be useful in assessing phenotype and guiding therapy.

The development of second generation antipsychotic agents, for example clozapine, olanzapine, quetiapine, amisulpride, aripiprazole and risperidone, has proved to be a significant advance in the treatment of schizophrenia. The atypical (or second-generation) antipsychotics have several therapeutic properties in common; however, they can differ significantly with regard to clinical potency and side-effects.

Therapeutic drug monitoring may assist in avoidance of extrapyramidal side-effects by maintaining minimum effective concentrations during chronic treatment, although for the majority of patients this is a matter of quality of life rather than safety. Monitoring is available from specialist centres in difficult cases. In the case of clozapine, there is a strong correlation between clozapine plasma concentrations and the incidence of seizures. Clozapine also carries a significant risk (~3%) of agranulocytosis and regular (weekly at the start of therapy) differential white blood cell monitoring is required.

Value of monitoring: moderate (haloperidol, clozapine) to low.

Antimicrobial drugs

Most antimicrobial drugs are well tolerated and do not require therapeutic drug monitoring. The exceptions to this general principle include the aminoglycoside antibiotics, the glycopeptides vancomycin and teicoplanin, and chloramphenicol.

Aminoglycosides (amikacin, gentamicin, tobramycin)

The aminoglycosides are an important group of antimicrobial drugs used in the treatment of severe systemic infection by some Gram-positive and many Gram-negative organisms. Amikacin, gentamicin and tobramycin are also active against *Pseudomonas aeruginosa*. Streptomycin is active against *Mycobacterium tuberculosis* and is now generally reserved for tuberculosis (see below). Neomycin is too toxic for systemic use and can only be used for topical application (skin, mucous membranes). The parent compounds are produced by moulds of the *Streptomyces* family (streptomycin, tobramycin, neomycin) or the *Micromonospora* family (gentamicin). The different host organisms account for the variation in spelling of the suffix.

The aminoglycoside antibiotics exhibit relatively simple pharmacokinetics. They are large, highly polar molecules with very poor oral bioavailability, and must be given parenterally. They are not protein bound and not metabolized, and are excreted through the kidneys. The plasma half-life is 2–3 h, except when renal function is impaired, but the drug may accumulate in tissues. If therapy is continued for longer than a week, tissue sites become saturated and plasma concentrations may rise.

These drugs exhibit significant systemic toxicity at plasma concentrations just above those necessary for bactericidal activity. The main toxic effects are nephrotoxicity and ototoxicity. Nephrotoxicity further reduces the ability to excrete aminoglycosides, and a vicious cycle may be precipitated. Nephrotoxicity is often reversible, as is mild ototoxicity, but severely damaged cochlear hair cells cannot be replaced and patients may be left with irreversible hearing loss and disturbed balance.

The drug concentration at which bactericidal effects are achieved (the minimum inhibitory concentration, MIC) is relatively easy to determine in vitro but often has little relevance in vivo, owing to variable penetrance of the drug to the site of infection and differing conditions at the infection site. Aminoglycosides also show a marked post-antibiotic effect – suppression of bacterial growth persists for some time after the drug is no longer present in the plasma. This makes definition of target plasma concentrations difficult, and this difficulty has been compounded in recent years by changes in the ways in which aminoglycosides are administered.

Until the mid-1990s, aminoglycosides were administered every 8 or 12 h, to give relatively stable plasma concentrations in view of their short half-life. It has become clearer that less frequent dosing (every 24 h or more) produces higher peak concentrations, which enhance bacterial kill, and lower trough concentrations, which reduce the systemic toxicity. Such regimens are more convenient,

less toxic, reduce adaptive resistance and are generally more suitable in patients with normal renal function. The approach was originally devised as once daily dosing, but a more accurate term is probably 'extended dosing interval', in which a plasma concentration measurement is used to design an individual dosing interval which reflects the patient's needs and renal function. The Hartford nomogram (Nicolau et al. 1995, see [Further reading](#)) is an example of this approach, but local guidelines should be consulted on dosage and serum concentrations. Monitoring is essential to achieve effective therapy while avoiding toxicity, particularly in infants, the elderly, the obese and patients with cystic fibrosis, if high doses are being used or if renal function is impaired.

Value of monitoring: high.

Glycopeptides (vancomycin and teicoplanin)

The glycopeptide antibiotic vancomycin is used intravenously in endocarditis and other serious infections caused by Gram-positive cocci including multi-resistant *Staphylococci* (MRSA). It is also used orally in the treatment of antibiotic-associated (pseudomembranous) colitis. Like the aminoglycosides, the glycopeptides are poorly absorbed, not metabolized, excreted renally and are potentially nephrotoxic and ototoxic. Indications for monitoring have been controversial, but there is definitely a role for monitoring in patients with poor renal function to achieve maximum effect with minimal toxicity. Teicoplanin is similar to vancomycin, but has a longer duration of action such that once-daily dosing is sufficient. No relationship between plasma concentration and toxicity has been established, and teicoplanin is not normally monitored.

Value of monitoring: moderate (vancomycin), low (teicoplanin).

Chloramphenicol

Chloramphenicol is a powerful broad-spectrum antibiotic that carries a risk of serious haematological side-effects when given systemically. It is used to treat life-threatening infections such as cholera, typhoid fever, resistant *Haemophilus influenzae*, septicaemia and meningitis. Chloramphenicol is metabolized in the liver to the active glucuronide, and peak plasma concentrations (2 h post-dose) in the range 10–25 mg/L (30–77 µmol/L) are desirable. Concentration monitoring is essential in neonates and recommended in children under four years, the elderly and patients with hepatic impairment.

Value of monitoring: moderate.

Antifungal drugs

The successful management of invasive fungal infections remains challenging to clinicians, and the incidence of invasive mycosis has risen in parallel with the rise in the population of immunocompromised patients. The morbidity and mortality caused by these infections remains

high, and TDM has a role in ensuring that adequate drug concentrations are attained at the site of infection without systemic toxicity. Four triazole antifungal drugs (fluconazole, itraconazole, posaconazole and voriconazole) have been approved for use. They show marked inter- and intraindividual variations in blood drug concentrations after dosing, owing to variable absorption from the gut (itraconazole, posaconazole) and polymorphism in the CYP2C19 enzyme (voriconazole). The pyrimidine analogue flucytosine also exhibits wide interpatient variation in blood concentration owing to variation in renal elimination. The high pharmacokinetic variability and need for optimal adjustment of drug exposure to ensure that infections are treated adequately are arguments in favour of TDM, but although concentration–effect relationships have been demonstrated for the triazoles and flucytosine, optimal target blood concentrations have not been definitively established and concentration measurements in clinically relevant timescales are not yet readily available. Itraconazole and flucytosine concentrations should be measured routinely in all patients during the first week of therapy and in patients with poor responses. Monitoring of posaconazole and voriconazole should be considered in patients who are not responding to therapy, those with gastrointestinal dysfunction, children and those taking drugs which interact with triazole antifungal agents.

Value of monitoring: moderate.

Antitubercular drugs

The treatment of active tuberculosis (TB) always requires the use of multiple antibacterial agents. It is usually treated in two phases – an initial phase using four drugs and a continuation phase using two drugs. The most frequently used regimen is isoniazid, rifampicin, pyrazinamide and ethambutol (or, rarely, streptomycin), followed by isoniazid and rifampicin. Isoniazid and rifampicin are the key components of this regimen, and both drugs show significant pharmacokinetic variability. The majority of patients are completely cured by standard regimens, and TDM has no role in dosage optimization in these patients. However, patients who are slow to respond to treatment, have drug-resistant TB, are at risk of drug–drug interactions or have concurrent disease states (e.g. acquired immunodeficiency syndrome) that significantly complicate the clinical picture may require individualization of drug therapy, and may benefit from TDM. Early intervention guided by concentration monitoring may prevent the development of further drug resistance. The pharmacokinetic interactions between antitubercular drugs and other medications can often be of considerable clinical concern. Rifampicin is a potent inducer of cytochrome P450 CYP3A and decreases plasma concentrations of the HIV-protease inhibitors significantly; isoniazid is a P450 enzyme inhibitor. Patients with HIV are at particular risk for drug–drug interactions.

When TDM is justified, the 2 h post-dose concentrations of isoniazid, rifampicin, pyrazinamide and ethambutol are usually most informative. Unfortunately, low 2 h values do not distinguish between delayed absorption or malabsorption and non-adherence. In such cases, a

second sample, often collected at 6 h post-dose, provides additional information. Trough concentrations of many anti-TB drugs are below the limit of detection of current assays and have limited relevance. Plasma concentrations required for effective therapy, especially on multi-drug regimens, are only partially known, as detailed pharmacokinetic and pharmacodynamic information in humans is lacking. There is a clear dose–response relationship for rifampicin and pyrazinamide and low isoniazid concentrations are also associated with poorer outcomes. Second-line TB drugs such as *p*-aminosalicylic acid, cycloserine and ethionamide are not usually monitored. Streptomycin is used against resistant organisms. It is nephrotoxic and should be used with great care in patients with renal impairment; concentration monitoring is essential in such patients.

Value of monitoring: moderate.

Antiretroviral drugs

Several classes of antiretroviral drugs are used in the management of infection by HIV:

- nucleoside reverse transcriptase inhibitors, e.g. zidovudine, abacavir, didanosine, emtricitabine, lamivudine, stavudine and tenofovir
- non-nucleoside reverse transcriptase inhibitors, e.g. efavirenz, etravirine and nevirapine
- protease inhibitors, e.g. atazanavir, darunavir, fosamprenavir, indinavir, lopinavir, nelfinavir, ritonavir, saquinavir and tipranavir
- fusion inhibitors, e.g. enfuvirtide
- entry inhibitors, e.g. maraviroc
- integrase inhibitors, e.g. raltegravir.

The nucleoside reverse transcriptase inhibitors (NRTIs) are pro-drugs that require activation by intracellular phosphorylation. Plasma NRTI concentrations therefore do not correlate with clinical effect and NRTIs are not suitable for TDM.

Non-nucleoside reverse transcriptase inhibitors (NNRTIs) display highly variable pharmacokinetics; the cytochrome P450 isoenzyme involved in their metabolism (CYP2B6) has a polymorphic variant resulting in slower metabolism and higher plasma concentrations. The protease inhibitors (PIs) also have highly variable pharmacokinetics, but their plasma concentrations have been shown to correlate with virological response. Therapeutic drug monitoring for NNRTIs and PIs may have an important adjunctive role in individualizing therapy in high-risk patients or those with PI resistance. Adherence remains a major problem, especially in developing countries where patient understanding is often poor. Concentrations are normally monitored 1 h post-dose. Protease inhibitors and NNRTIs can be quantitated by liquid chromatography tandem mass spectrometry (LC-MS) in a single run.

Value of monitoring: low to moderate.

Antineoplastic drugs

Despite their considerable toxicity, there are no controlled trials that indicate clear advantages to routine

concentration monitoring of anticancer drugs, with the exception of methotrexate (see below). Antineoplastic agents are often used in combinations of up to four drugs, which makes the establishment of therapeutic ranges difficult. Areas under the concentration–time curve give a better indication of total exposure to the drug than isolated plasma concentrations, but are more difficult to apply routinely. Pharmacokinetic optimization protocols for many classes of cytotoxic compounds have been used in some specialized centres and a role for monitoring has been proposed for some newer drugs such as the tyrosine kinase inhibitor imatinib in chronic myeloid leukaemia.

Value of monitoring: low (except methotrexate – see below).

Methotrexate

Methotrexate is widely used in a variety of chemotherapeutic regimens for malignancies, where it is given in high doses by the intravenous route over periods of 6–12 h. It is also used orally at low dose (7.5–25 mg weekly) in Crohn disease, rheumatoid arthritis and as an antipsoriatic agent.

Methotrexate is an inhibitor of folic acid metabolism. The rationale behind high-dose therapy is that a short period of exposure to high concentrations of drug will eliminate rapidly dividing cells, while sparing those growing more slowly. At the end of the infusion, methotrexate concentrations should fall rapidly. Where this is not the case, or where the rate of infusion produces too high a plasma concentration, folinic acid (leucovorin) is infused to provide a source of folate and reverse toxic effects. Measurement of methotrexate concentrations may therefore be required in two circumstances – to allow adjustment of infusion to maintain a steady plasma concentration at 10^{-4} M and to ensure that the concentrations at 24 and 48 h after the start of the infusion have fallen to <5 – 10 $\mu\text{mol/L}$ and <0.5 – 1.0 $\mu\text{mol/L}$, respectively, depending on the protocol. A rapid turnaround time is required. Folinic acid may interfere in some assay systems, complicating measurements made after treatment has been given. Methotrexate is metabolized to 7-hydroxymethotrexate and both the parent drug and metabolite are excreted via the kidneys and are nephrotoxic due to their low solubility; prehydration, alkalization and good renal function are essential for the safe use of this drug. Hepatotoxicity has been reported in a number of patients. Monitoring of liver function and the plasma concentration of the N-terminal propeptide of collagen type III to detect hepatic fibrosis is advised in patients on long-term oral treatment; renal function and white blood cell counts should also be monitored. Methotrexate concentration monitoring is not usually necessary in patients on low-dose regimens.

Value of monitoring: high (on high-dose regimens).

Bronchodilator drugs

Theophylline/caffeine

Theophylline is a bronchodilator used for the relief of asthma that has a number of other pharmacological effects, notably the stimulation of cardiac output and the

dilatation of peripheral blood vessels. It is no longer a first- or second-line drug in asthma, but is still useful in patients who have difficulty with inhalers and those with predominantly nocturnal symptoms. Theophylline has a short half-life (3–13 h in adults), but the availability of sustained-release preparations compensates for this and allows dosing at 12 or 24 h intervals. It also has a narrow therapeutic window and wide pharmacokinetic variability, so it is important to have access to TDM if the drug is to be used effectively and safely.

In outpatients, adherence has been shown to be erratic, with as few as 30% of patients taking the drug as prescribed. In such cases, TDM can contribute little unless the patient is toxic or grossly non-compliant. However, in inpatients, and especially those receiving intravenous therapy, TDM can be of great assistance in dosage optimization and in confirming toxicity. Steady-state trough concentrations are the most reproducible for monitoring, except when toxicity is suspected.

Theophylline is metabolized in the liver by a variety of pathways, including oxidation via the cytochrome P450 system (CYP1A2/CYP2E1), which is subject to induction by drugs such as phenytoin. In addition to this, theophylline shows dose-dependent excretion. It is therefore preferable that those involved in its monitoring are familiar with the pharmacokinetic optimization programmes that are available.

The target range in adults is 10–20 mg/L (55–110 $\mu\text{mol/L}$). The range is lower in neonates (5–15 mg/L; 25–80 $\mu\text{mol/L}$). This is partly due to sensitivity and partly because babies under six months old metabolize theophylline in a different manner to adults, converting a proportion to caffeine, an active metabolite. After six months, caffeine production is replaced by the adult pattern of metabolism to a variety of methylated urates, including 3-methylxanthine, which is also active. Caffeine has now taken over as the drug of choice to treat neonatal apnoea, as dose regimens are simpler and the effects are more predictable. Caffeine concentrations in the range 5–20 mg/L (25–100 $\mu\text{mol/L}$) are generally associated with response in neonatal apnoea, but the lower toxicity and simpler pharmacokinetics mean that TDM is rarely required.

Side-effects of theophylline are relatively frequent – mild or moderate effects such as nausea, headache and jitteriness are common within the target range, and more serious effects (tremor, agitation, insomnia, diarrhoea, palpitations, seizures and cardiac arrhythmias) occur with increasing frequency at plasma concentrations >20 mg/L (110 $\mu\text{mol/L}$). Caffeine produces much less tachycardia and fewer fits.

Oral therapy with theophylline may involve the free base or any of a number of salts. For intravenous therapy, a water-soluble salt such as aminophylline (theophylline ethylenediamine) is used (aminophylline is approximately 80% w/w theophylline).

Value of monitoring: high (theophylline); low to moderate (caffeine).

Immunosuppressants

Monitoring of immunosuppressants has become an essential adjunct to drug therapy for organ transplant

patients because of the relatively narrow therapeutic index of these agents. Therapeutic drug monitoring is important for three of the major immunosuppressants in current use: ciclosporin, tacrolimus, and sirolimus. All three of these drugs are lipophilic and are concentrated in red cells, so whole blood samples are commonly used for analysis. The indications for and approach to monitoring of mycophenolic acid are more controversial.

Ciclosporin

Ciclosporin is a cyclic peptide, widely used for prevention of graft rejection following kidney, liver, heart, lung, heart–lung, bone marrow or pancreas transplants. It is also used in lower doses as an immunosuppressant in severe psoriasis, atopic dermatitis, nephrotic syndrome, ulcerative colitis and rheumatoid arthritis, and in the prophylaxis and treatment of graft-versus-host disease. Ciclosporin inhibits calcineurin phosphatase and limits T cell activation, producing immunosuppression without myelosuppression. However, it is nephrotoxic, and the major difficulty with its use lies in balancing the risks of undertreatment, leading to rejection of the organ, and overtreatment, which can cause kidney damage. This is particularly relevant in renal transplantation, where it is essential to determine whether a rising plasma creatinine concentration is due to nephrotoxicity or rejection.

The pharmacokinetics of ciclosporin vary widely between patients, and target plasma concentrations are dependent on the transplant type, time after transplantation, sample time post-dose, analytical method used and other drugs administered. It is not possible to give specific guidance here and local protocols should be consulted. The drug is highly lipophilic and accumulates in red cells, so EDTA whole blood is the preferred matrix for assay. Trough (pre-dose) concentrations have traditionally been used; although concentrations at 2 h post-dose (C₂) give a better indication of ciclosporin exposure, they do require precise timing. Immunoassays and LC-MS assays have been used. Ciclosporin is metabolized in the liver by CYP3A, producing more than 30 metabolites that have variable cross-reactivity in immunoassays, explaining some of the differences in apparent ciclosporin concentrations determined by different assays.

Value of monitoring: high.

Sirolimus

Sirolimus is a macrolide antibiotic, also known as rapamycin. It is structurally similar to tacrolimus, but has a different mode of action. Both sirolimus and tacrolimus bind to a binding protein (FK binding protein, FKBP), but the sirolimus-FKBP complex then binds to a protein kinase, mammalian target of rapamycin (mTOR). This interaction blocks signal transduction in the second phase of T cell activation. Sirolimus reduces acute rejection when co-administered with a calcineurin phosphatase inhibitor and steroids, and allows earlier withdrawal of ciclosporin. When sirolimus is given concurrently with the microemulsion formulation of ciclosporin, the absorption

of sirolimus is markedly enhanced, almost doubling the effective dose, so it is recommended to give sirolimus 4 h after ciclosporin. The drug is metabolized in the liver by CYP3A, and is also a substrate for the P-glycoprotein transporter system. Sirolimus is not nephrotoxic so is particularly useful in kidney transplantation, but has a range of other adverse effects, including lipid disturbances, anaemia, leukopenia and thrombocytopenia and impaired wound healing. Over-suppression of the immune system is associated with infection and neoplasia. Blood concentrations are a better guide to effect than dose is, and concentration monitoring is advised, particularly in patients with hepatic impairment or on interacting drugs. Pre-dose (trough) concentrations measured on EDTA whole blood should be in the region of 4–12 µg/L when the drug is being given with ciclosporin and 12–20 µg/L when ciclosporin has been withdrawn. Liquid chromatography or LC-MS are the preferred analytical methods, but an immunoassay is also available.

Value of monitoring: high.

Tacrolimus

Like ciclosporin, tacrolimus is a calcineurin phosphatase inhibitor used to prevent or treat graft rejection in kidney, liver and heart transplant patients. It is also metabolized in the liver by CYP3A and is concentrated in red blood cells, so EDTA whole blood samples are used for monitoring. Unlike ciclosporin, it is not intrinsically nephrotoxic, but does cause renal vasoconstriction, which leads to chronic graft dysfunction. Adverse effects are minimized by using the lowest dose consistent with efficacy and by the use of combined therapy with other drugs such as mycophenolic acid, allowing dose minimization for all drugs. Monitoring is essential, particularly in hepatic impairment and during or following treatment with drugs that induce or inhibit metabolism. Target concentration ranges vary with transplant type, other drug therapy and analytical method, but are typically in the range 5–15 µg/L. Immunoassays or LC-MS are used.

Value of monitoring: moderate to high.

Mycophenolic acid

Mycophenolic acid (MPA) is increasingly used in solid organ transplantation as an antiproliferative immunosuppressant. It is used for the prophylaxis of acute rejection in kidney, liver and heart transplantation in combination with calcineurin phosphatase inhibitors (ciclosporin/tacrolimus) and corticosteroids. It is given as the pro-drug mycophenolate mofetil (MMF), or as the enteric-coated sodium salt of MPA. It is rapidly absorbed and mainly excreted renally, with <10% hepatic metabolism to the inactive glucuronide. Concentrations in plasma are relatively high (1–3 mg/L; 3–9 µmol/L) and trough plasma (EDTA or heparin) samples are often used for monitoring, although the optimum measures of MPA exposure are still the subject of considerable debate as trough plasma concentrations vary widely. Limited-sampling strategies to

give a better estimate of the area under the concentration–time curve have been advocated. The general routine use of MPA monitoring in solid-organ transplantation is not currently recommended, though monitoring has a role in specific patient populations, including those at increased immunological risk or those with changing renal or hepatic function.

Value of monitoring: low to moderate.

Opiate and opioid drugs

Methadone/buprenorphine

Methadone and buprenorphine are synthetic opioids used as adjunct therapy in the treatment of opiate dependence. The primary requirement in this setting is to confirm that patients have been taking their medication, rather than selling it, since these drugs have a street value. Qualitative testing of urine is used to achieve this. For patients on methadone therapy, urine is often tested for the methadone metabolite EDDP (2-ethylidine-1,5-dimethyl-3,3-diphenyl pyrrolidine), as this guards against adulteration of urine specimens with the medication and confirms that drug has actually been ingested. Monitoring of plasma concentrations has not yet been shown to be routinely useful, despite significant pharmacokinetic variation and much recent interest in dosage optimization for opiate substitutes.

Value of monitoring: low.

Morphine

Morphine is used as a short-term analgesic, often as patient-controlled analgesia following surgical procedures, and for long-term analgesia in conditions with protracted pain, for example the later stages of malignant disease. Morphine kinetics are complex and absorption can be erratic. There is also an active metabolite, morphine 6-glucuronide (M6G), with analgesic properties superior to morphine. Heroin (diacetylmorphine) is converted rapidly to morphine via a short-lived monoacetylmorphine intermediate. Until recently, morphine was not part of the repertoire of drugs for which therapeutic monitoring was advocated, but there is now evidence that, in patients in whom analgesia is difficult to achieve, plasma morphine concentrations can be useful. The method of choice is high performance liquid chromatography – mass spectrometry, since the active metabolites M6G and codeine can also be determined; this is of particular use in patients with renal insufficiency in whom there may be accumulation of M6G.

Value of monitoring: low to moderate.

ACKNOWLEDGEMENT

I am grateful to Dr Michael J. Stewart who was the author of this chapter in previous editions of the book.

Further reading

General pharmacokinetics

Brunton L, Chabner BA, Knollman B, editors. Goodman and Gilman's The pharmacological basis of therapeutics. 12th ed. New York: McGraw-Hill; 2011.

A huge compendium of knowledge of all aspects of therapy. Regularly updated and exceptionally valuable. Appendices give the key pharmacokinetic data on most drugs of interest and useful clinical data on toxicity.

Clark W, McMillin G. Application of TDM, pharmacogenomics and biomarkers for neurological disease pharmacotherapy: focus on antiepileptic drugs. *Pers Med* 2006;3:139–49.

Hallworth M, Watson I. Therapeutic drug monitoring and laboratory medicine. London: ACB Venture Publications; 2008.

A useful practical guide for clinical laboratories.

Hammett-Stabler CA, Dasgupta A. Therapeutic drug monitoring data – a concise guide. Washington DC: AACCC Press; 2007.

Invaluable concise guide.

Llorente Fernandez E, Pares A, Ajuria I et al. State of the art in therapeutic drug monitoring. *Clin Chem Lab Med* 2010;48:437–46.

Review of the application of quality management to TDM.

Macleod S. Therapeutic drug monitoring in pediatrics: how do children differ? *Ther Drug Monit* 2010;32:250–2.

Rowland M, Tozer TN. Clinical pharmacokinetics and pharmacodynamics – concepts and applications. Baltimore, MD: Lippincott, Williams & Wilkins; 2011.

Excellent and long-established general pharmacokinetic text.

Analytical methods

Adaway JE, Keevil BG. Therapeutic drug monitoring and LC-MS/MS. *J Chromatogr B Analyt Technol Biomed Life Sci* 2012;883–884:33–49.

Brandhorst G, Oellerich M, Maine G et al. Liquid chromatography-tandem mass spectrometry or automated immunoassays: what are the future trends in therapeutic drug monitoring? *Clin Chem* 2012;58:821–5.

Hallworth M. Therapeutic drug monitoring. In: Moffat AC, Osselson D, Widdop B et al. editors. Clarke's analysis of drugs and poisons. 4th ed London: Pharmaceutical Press; 2011. p. 59–72.

Contains comprehensive tables with suggested analytical approaches for commonly monitored drugs.

Antiarrhythmics/cardiac glycosides

Dasgupta A. Therapeutic drug monitoring of digoxin: impact of endogenous digoxin and exogenous digoxin-like immunoreactive substances. *Toxicol Rev* 2006;25:273–81.

A useful explanation of this intriguing analytical problem.

Anticonvulsants

Patsalos PN, Berry DJ, Bourgeois BFD et al. Antiepileptic drugs – best practice guidelines for therapeutic drug monitoring: A position paper by the sub-commission on therapeutic drug monitoring, I L A E Commission on therapeutic strategies. *Epilepsia* 2008;49:1239–76.

Authoritative review.

Antidepressants/antipsychotics

Hiemke C, Baumann P, Bergemann N et al. AGNP consensus guidelines for therapeutic Drug monitoring in psychiatry: update 2011. *Pharmacopsychiatry* 2011;44:195–235.

Comprehensive review.

Antimicrobial drugs

Dasgupta A. Advances in antibiotic measurement. *Adv Clin Chem* 2012;56:75–104.

Nicolau DP, Freeman CD, Belliveau PP et al. Experience with a once-daily aminoglycoside program administered to 2,184 patients. *Antimicrob Agents Chemother* 1995;29:650–5.

Antifungal drugs

Hussaini T, Ruping MJ, Farowski F et al. Therapeutic drug monitoring of voriconazole and posaconazole. *Pharmacotherapy* 2011;31:214–25.

Antitubercular drugs

Magis-Escurra C, van den Boogard J, Ijdema D et al. Therapeutic drug monitoring in the treatment of tuberculosis patients. *Pulm Pharmacol Ther* 2012;25:83–6.

Antiretrovirals

Pretorius E, Klinker H, Rosenkrantz B. The role of therapeutic drug monitoring in the management of patients with human immunodeficiency virus infection. *Ther Drug Monit* 2011;33:265–74.

Antineoplastic drugs

Bach DM, Straseski JA, Clarke W. Therapeutic drug monitoring in cancer chemotherapy. *Bioanalysis* 2010;2:863–79.

Immunosuppressants

Brandhorst G, Oellerich M, Brunet M et al. Individually tailored immunosuppression – is there a role for biomarkers? *Clin Chem* 2011;57:376–81.

Taylor PJ, Tai CH, Franklin ME et al. The current role of liquid chromatography-tandem mass spectrometry in therapeutic drug monitoring of immunosuppressant and retroviral drugs. *Clin Biochem* 2011;44:14–20.

Wavamunno MD, Chapman JR. Individualization of immunosuppression: concepts and rationale. *Curr Opin Organ Transplant* 2008;13:604–8.

Methadone/buprenorphine

Brunen S, Vincent PD, Baumann P et al. Therapeutic drug monitoring for drugs used in the treatment of substance-related disorders: literature review using a therapeutic drug monitoring appropriateness rating scale. *Ther Drug Monit* 2011;33:561–72.

APPENDIX 39.1: CALCULATIONS FOR THE DETERMINATION OF DOSE REQUIREMENTS TO ACHIEVE STEADY-STATE CONCENTRATIONS

The loading dose depends on the volume of distribution (V_d). Hence, to calculate a loading dose use:

$$\text{Loading dose} = \frac{V_d \times \text{desired plasma concentration } (C_p)}{\text{Bioavailability } (F) \times \text{correction factor for formulation } (S)}$$

The maintenance dose required for steady-state depends on the clearance and is calculated by:

$$\text{Maintenance dose} = \frac{\text{Clearance } (Cl) \times \text{desired plasma concentration } (C_p) \times \text{interval between doses } (T)}{\text{Bioavailability } (F) \times \text{correction factor for formulation } (S)}$$

Average population values for clearance, volume of distribution, bioavailability and formulation correction factor (if relevant) may be found in reference texts.

Poisoning

James W. Dear

CHAPTER OUTLINE

INTRODUCTION 787**AETIOLOGY OF POISONING 788**

Intrauterine 788

Neonates 788

Infants 788

Childhood 788

Adult life 788

TYPES OF LESION IN POISONING 788**DIAGNOSIS AND MANAGEMENT OF POISONING:****GENERAL PRINCIPLES 789**

Diagnosis 789

Management 791

SPECIFIC POISONS 795

Paracetamol (acetaminophen) 795

Salicylate 796

Chloroquine 798

Digoxin 798

Iron 798

Other metals 798

Organophosphates 799

Alcohols and glycols 799

Drug and substance abuse 800

Benzodiazepines 803

Theophylline 803

Antidepressants 803

Lithium 804

Cyanide 805

Carbon monoxide 805

Methaemoglobinaemia 805

Plant and fungal toxins 806

CONCLUSION 807**APPENDIX 807****INTRODUCTION**

Poisoning can be defined as an interaction between a foreign chemical (toxin) and a biological system that results in damage to a living organism. In general, the medical profession has been more concerned with the acute effects of toxins and the clinical management of toxicity, but chronic effects of toxins have much more importance on a global scale. For example, acute ingestion of ethanol may result in intoxication, which can cause death directly from its acute depressant effects and also through intoxication-related accidents and violence. However, alcohol-related liver disease, myopathy, hypertriglyceridaemia and cancers are all effects of chronic alcohol abuse and together, cause far more deaths than the acute effects of alcohol. Other types of chronic or delayed effects of toxins include mutagenicity, carcinogenicity and teratogenicity. For example, many cases of bronchogenic carcinoma are a direct result of exposure to cigarette smoke. Increasing concerns about environmental contamination by chemicals has added a further dimension to toxicology, especially as many chemicals, such as pesticide residues, may be found in small amounts in the general population, who do not appear to be suffering any ill effects.

The interaction between a toxin and the biological system is highly complex. Once the biological system has

been exposed to the toxin, the toxin is subject to various processes. In the case of a drug, this is usually termed pharmacokinetics, but in the case of a poison, the term toxicokinetics may be used. This includes absorption, for which the route of administration is of great importance, as are the dose of the toxin and the duration of exposure. Once a toxin has entered the bloodstream, its distribution depends to a large extent, on its physicochemical characteristics, including its degree of lipid or water solubility and its degree of ionization. Metabolism is a further important process, which usually renders toxins inactive, but which may also convert some substances of low toxicity into highly toxic metabolites. Classic examples of this include paracetamol (acetaminophen), methanol and ethylene glycol. Finally, toxins are excreted, although some remain within the body for many years. All of these processes determine the amount of toxin that interacts with tissues to produce a toxic effect.

The study of the interaction between toxins and tissues is called toxicodynamics. An understanding of the nature of this interaction is essential for an intelligent approach to the diagnosis and management of poisoning. This interaction may cause functional and metabolic disturbances that produce the characteristic clinical features of poisoning and the laboratory changes that may be essential for diagnosis or management. These changes are summarized in [Figure 40.1](#).

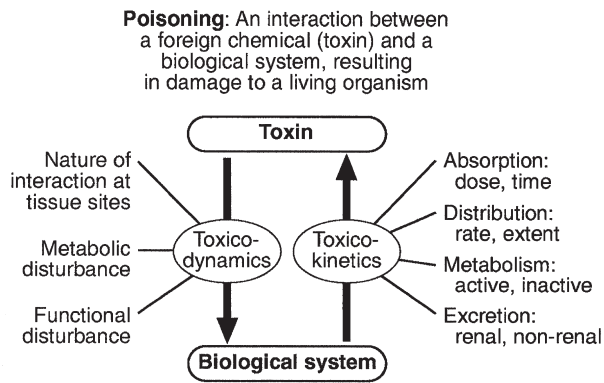


FIGURE 40.1 ■ The factors involved in poisoning.

This chapter covers several of the more common types of acute poisoning and also some aspects of chronic poisoning. Clinical features are given in some detail, as these frequently integrate with the metabolic and biochemical changes.

AETIOLOGY OF POISONING

Intrauterine

A number of drugs, such as tretinoin, should not be taken by women intending to become pregnant because of their known teratogenicity. Other drugs, such as phenytoin and several other anticonvulsants, are teratogenic but their continued use may be considered necessary during pregnancy. As far as poisoning and overdose during pregnancy are concerned, the fetus generally only suffers damage secondary to the mother's illness, and the management is no different from that of overdose in non-pregnant women. One exception is carbon monoxide, which has a higher affinity for fetal haemoglobin than adult haemoglobin; the mother should be treated aggressively even in apparently mild poisoning. Antidotes such as acetylcysteine or desferrioxamine should not be withheld because the patient is pregnant.

Neonates

During the neonatal period, particularly in premature infants, there is a risk of iatrogenic overdose because of the poor metabolic and excretory capacity relative to body weight. Particular care should be taken with drugs such as theophylline, digoxin, chloramphenicol and morphine.

Infants

The commonest cause of poisoning between the ages of one and four is termed accidental poisoning because of the natural exploratory activities of children at this stage of life. Each year, this results in 20 000 admissions to hospital in Britain, but only about ten fatalities. The commonest agents involved are paracetamol and oral contraceptives, but most fatalities occur from tricyclic antidepressants, salicylates, iron, methadone and quinine. Another form of poisoning in this age group is non-accidental injury,

which is usually inflicted by a carer such as the mother, and the child may present with repeated episodes of floppiness and collapse, depending on the substance responsible. This is most commonly with a drug prescribed for the mother, though salt and other readily available chemicals have been used (see Chapter 44).

Childhood

Later in childhood, substance abuse, particularly volatile substance abuse ('glue sniffing'), can cause serious problems. There are about 30 deaths each year in Britain from this cause. Acute and chronic alcohol toxicity are often underestimated in this age group. Intentional drug overdose is common, but suicide is rare.

Adult life

In early adult life, particularly in females, drug overdose in the form of parasuicidal gestures is common, though rarely fatal. In later adult life, suicidal intent is commoner and parasuicidal gestures are less frequent. The commonest causes of suicidal poisoning in England are tricyclic antidepressants, analgesic drugs and carbon monoxide. Male suicides are approximately three times as frequent as female suicides. Accidental poisoning in adults is usually domestic, the commonest agent being carbon monoxide, or industrial, in which a wide range of chemicals may be involved. Homicidal poisoning, though rare, may go undetected if it is not suspected.

TYPES OF LESION IN POISONING

Many types of poisoning involve a highly specific interaction between the toxin and one type of tissue. This has led to a 'target organ' oriented approach in toxicology, which is of great help in understanding poisoning, where damage to a single organ or tissue produces a characteristic clinical and pathological picture. Examples include poisoning with paracetamol, salicylates, cholinesterase inhibitors and cardiac glycosides. In other instances, the type of toxicity may be less clearly definable. The acute effects of ethanol are mainly due to its central nervous depressant effect. Cocaine, tricyclic antidepressants, many volatile substances, some β -adrenergic blocking drugs and dextropropoxyphene can all produce cardiotoxicity due to a quinidine-like sodium channel blocking effect. However, the clinical features are definable as a toxic effect on the heart.

An important way in which many poisons can affect the organism is by interference with the oxygen pathway from the inspired air to cellular respiration. Thus, a reduction in the oxygen content of the inspired air may cause acute hypoxaemia, rapidly leading to collapse and coma. Interference with the mechanics of respiration by poisons may result in hypoxaemia and hypercapnia (type II respiratory failure), while disturbances of oxygen transfer produce hypoxaemia alone (type I respiratory failure). Even if the oxygen content of the air is not diminished and respiration is functioning adequately,

disturbances to other processes may prevent oxygen reaching its site of action (see Chapter 5). The oxygen-carrying capacity of the blood may be reduced by the presence of carboxyhaemoglobin or methaemoglobin or, more rarely, by acute haemolysis. Cardiac output may be reduced by a number of poisons that cause cardiac arrhythmias, depress the contractility of the heart or cause extreme vasodilatation. The final step where poisons may interfere with the oxygen pathway is blockage of the cytochrome enzyme chain as occurs with toxins such as cyanide and hydrogen sulphide. These effects and the main poisons involved are summarized and outlined in Table 40.1.

DIAGNOSIS AND MANAGEMENT OF POISONING: GENERAL PRINCIPLES

Diagnosis

In most cases of poisoning, the diagnosis is provided by the history and clinical examination. In some patients, however, there may be no history, though poisoning may be suspected because of the circumstances or the clinical features. In rare cases, the clinical presentation may mimic other medical conditions and it is necessary to consider the diagnosis of poisoning, especially when there are unusual features of the illness. Examples of these conditions include carbon monoxide, lead and paraquat poisoning, and drug and substance abuse.

The clinical features of poisoning depend most of all on the type of agent involved, but also on the route of exposure (oral, intravenous, percutaneous or inhaled), the duration of exposure and time elapsed since exposure, and also the patient's age and medical background

(conditions such as diabetes, asthma, chronic renal failure and epilepsy). Clinical examination may reveal signs of injection marks, cutaneous burns, buccal corrosion, tablet residues in the mouth or cutaneous blisters. These blisters (originally called 'barbiturate blisters') may be found following overdose with barbiturates, benzodiazepines, tricyclic antidepressants or chloral hydrate, or exposure to carbon monoxide. There may be evidence of pulmonary aspiration or secondary hypostatic pneumonia in comatose patients.

Clinical examination may indicate that a specific agonist/receptor pathway has been either over-stimulated or blocked by a poison. These syndromes are termed toxidromes and provide a useful initial guide to poison identification when the agent involved is unknown (Table 40.2).

In the severely ill patient, baseline haematological and biochemical investigations will be necessary and arterial blood gas estimations will also be required. A chest radiograph is usually indicated. The results of haematological and biochemical investigations may be of great assistance in making the diagnosis and monitoring the response to treatment. The tests to be carried out and their frequency depend very much on the poison involved. Many of these are given later in this chapter, and a general summary is given in Table 40.3. Toxicological investigations may also be required: in some cases, these may only be available in specialized laboratories. The usual samples required are blood (note that whole blood is required for measurement of carboxyhaemoglobin, methaemoglobin, cyanide and several metals) and urine (50 mL in a universal container with no preservative). Urine should always be provided where possible, since many types of analysis, such as drugs of abuse, are more efficiently carried out on urine samples. Some of these assays are qualitative,

TABLE 40.1 Some typical causes of blockade of the oxygen pathway due to poisoning, with mechanisms and typical values

Note that whatever the initial effect of a poison, there may be subsequent progression down the list of effects towards cell death. The effects of two or more poisons may be additive

Cause	Mechanism	Effect
Asphyxiant gases (e.g. butane, methane, carbon dioxide, nitrogen)	Hypoxic gas mixture	Reduced inspired oxygen fraction (F_{iO_2}) (normal ~0.21)
Respiratory depression (e.g. opioids, barbiturates, other sedatives and hypnotics). Respiratory muscle disorders: paralysis (e.g. organophosphates, botulinum toxin) or spasm (e.g. strychnine, phencyclidine)	Failure of ventilation (type II respiratory failure)	Reduced alveolar oxygen tension (P_AO_2) (normal ~13.3 kPa)
Aspiration pneumonitis Adult respiratory distress syndrome (e.g. paraquat)	Failure of oxygen transfer (type I respiratory failure)	Reduced arterial oxygen tension (P_aO_2) (normal 10–13.3 kPa)
Carboxyhaemoglobin (carbon monoxide); methaemoglobin (e.g. nitrites); haemolysis (e.g. arsine, stibine)	Loss of functioning haemoglobin	Reduced arterial oxygen content (CaO_2) (normal 18–21 volumes %)
Myocardial depressants (e.g. β -blockers, calcium antagonists, tricyclic antidepressants, dextropropoxyphene, cocaine)	Reduced cardiac output	Reduced tissue oxygen delivery (QO_2) (normal 12–16 mL/kg per min)
Chemical asphyxiants (cyanide, hydrogen sulphide)	Block of cytochrome enzyme chain	Reduced tissue oxygen consumption (VO_2) causing failure of oxidative metabolism (normal 3–4 mL/kg per min). Cell death

TABLE 40.2 Syndromes caused by poisoning

Clinical features	Example agent(s)
Anticholinergic syndrome (muscarinic antagonist) Dry skin Hyperthermia Thirst Dry mouth Dilated pupils Tachycardia Urinary retention Decreased bowel sounds	Tricyclic antidepressants (TCA) Antipsychotics Antihistamines <i>Atropa belladonna</i> Inocybe mushrooms
Antimitotic syndrome (cytotoxic to dividing cells) Bone marrow suppression (aplastic anaemia, leucopenia, thrombocytopenia) Alopecia Vomiting, diarrhoea, mucositis	Antineoplastic drugs Colchicine Radiation exposure
Cholinergic syndrome (nicotinic and muscarinic agonist) Defaecation Urination Miosis Bradycardia Tachycardia Emesis Lacrimation Hyperhidrosis Muscle paralysis and respiratory failure	Organophosphates Carbamate insecticides Nicotine Laburnum species Hemlock species
Corrosive poisoning GI tract pain Vomiting Haematemesis Dyspnoea Drooling Aspiration/inhalation causing stridor/pneumonitis Skin contact causing pain, ulceration, necrosis Eye contact causing inflammation of all layers	Hydrofluoric acid Sodium hydroxide Paraquat
Fume fever History of unpleasant smells Chills Cough Dyspnoea Headache Myalgia Malaise	Metal oxides (esp zinc oxide) Polymer fumes Other toxic industrial chemicals
Metabolic acidosis Deep and rapid (Kussmaul) breathing Obtunded consciousness Tachycardia Hypotension	Ethylene glycol Methanol Aspirin Paracetamol Iron Cyanide Carbon monoxide
Methaemoglobinaemia Blue-grey 'apparent' central cyanosis (blue to grey lips, tongue and mucus membranes, and slate grey skin) Persistent cyanosis despite oxygenation Fatigue, dizziness, headaches Depressed consciousness Seizures Urine may be discoloured black or brown	Benzene derivatives (phenols, cresols, aniline) Sodium nitrite Organic nitrites Chlorates Copper salts Prilocaine
Opioid syndrome Depressed consciousness Hypoventilation Pinpoint pupils Naloxone response Hypotension Pulmonary oedema	Opioids

(Continued)

TABLE 40.2 Syndromes caused by poisoning (Continued)

Clinical features	Example agent(s)
Sedative – hypnotic poisoning Depressed consciousness Ataxia Dysarthria Nystagmus	Ethanol Benzodiazepines and related drugs Gamma hydroxybutyrate (GHB) Gamma butyrolactone (GBL) Barbiturates
Serotonin agonist syndrome Restlessness Agitation Confusion Hyperreflexia Clonus Tremor Shivering Hypertonia Fever/hyperthermia Flushing Seizures	Serotonin-specific re-uptake inhibitors (SSRI) Monoamine oxidase Inhibitors (MAOI) Tricyclic antidepressants Venlafaxine Methylenedioxymetamphetamine (MDMA) Amfetamines Cocaine Tramadol Triptans Linezolid St John's wort 'Legal highs' Psilocybe mushrooms
Sympathomimetic syndrome Hyper/hypotension Tachycardia Neurological excitation Tremor Hyperreflexia Seizures	Cocaine Amfetamines 'Legal highs'
Vesicant poisoning (blistering agents) Conjunctivitis Keratitis Dermatitis Severe blistering	Nitrogen mustards Methyl bromide

but may be necessary in order to confirm the presence of compounds, which are not easily detectable in blood.

Poisons centres now exist in most countries, with specialized facilities for provision of information on the toxicity of substances and on the management of poisoning. Most hold stocks of specialized antidotes, such as snake antivenoms. Many also have an analytical laboratory and some have patient care facilities. They are usually accessible by telephone or e-mail. In the UK, there are several poisons centres that provide information on poisoning; see [Appendix 40.1](#).

Management

When an acutely ill patient presents, the first priority, regardless of the cause of the illness, is to establish the airway and provide immediate supportive treatment. Cardiopulmonary resuscitation or urgent ventilatory support may be needed. Depending on the patient's state, questioning about the possible agent involved, obtaining an antidote or taking diagnostic samples may waste valuable time. In view of the frequency of opioid abuse, naloxone (an opioid antagonist) may have both a diagnostic and a therapeutic role. Once any immediate life-threatening problems have been dealt with, full supportive care must

be provided while attempts are made to establish the diagnosis and treat the poisoning.

Respiratory support

In comatose patients, the most important measure is to maintain the airway and support ventilation as required, since respiratory complications are the commonest causes of death in unconscious poisoned patients. The unconscious patient should be placed in the left lateral (recovery) position in order to keep the airway patent and to minimize the risk of aspiration of gastric contents. Regular observation is essential as mechanical ventilation is not infrequently required.

Cardiovascular support

The circulation must be supported in order to maintain tissue perfusion and the electrocardiogram should be monitored, since cardiac dysrhythmias are common in poisoning. Antiarrhythmic agents and inotropic agents may be necessary.

Central nervous system complications

A brief convulsion due to cerebral hypoxia or to a toxic effect of the poison is not an indication for anticonvulsant

TABLE 40.3 Indications for measurement of drugs and poisons or other biochemical tests in actual or suspected poisoning

Toxin or measurement	Indications for measurement	Interpretation
Amfetamines (including MDA and MDMA)	Abuse	Tests can confirm exposure
Carbamazepine	1. Overdose	Ataxia >12 mg/L (50 µmol/L) Severe poisoning >40 mg/L (170 µmol/L)
Carboxyhaemoglobin	2. Therapeutic monitoring	Therapeutic range 5–10 mg/L (20–40 µmol/L)
	1. Carbon monoxide inhalation	May confirm exposure and indicate severity (normal <1%; <5% in smokers)
	2. Smoke-exposed fire victims	Carboxyhaemoglobin >12% indicates potential for pulmonary damage
Chloroquine	3. Methylene chloride exposure	Carbon monoxide is a metabolite of methylene chloride (dichloromethane)
	Accidental or deliberate overdose	May confirm ingestion and indicate potential toxicity
Cholinesterase (red blood cell, RBC)	1. Acute poisoning with organophosphates	RBC cholinesterase <20% of normal indicates significant exposure (usually by ingestion)
	2. Chronic exposure to organophosphates	Plasma pseudocholinesterase <50% of normal helps to confirm exposure but is a poor guide to severity
Digoxin	1. Digoxin or digitoxin overdose	In patients with severe poisoning blood digoxin concentration >6 h after overdose can be used to guide the dose of digoxin specific antibody
	2. Detecting digitalis immunoreactivity in cardiac glycoside plant ingestion	May be used to confirm ingestion but is not a guide to severity
	3. Detecting digitalis toxicity in therapeutic use	To confirm toxicity or undertreatment. Hypokalaemia may cause signs of toxicity with digoxin concentrations in the therapeutic range
Drug abuse screen	1. Suspected acute toxicity	May confirm diagnosis or need for treatment
	2. Medicolegal indications	May be used to confirm abstinence (sample collection must be supervised)
Ethanol	3. Employment screening	May exclude from certain types of employment
	1. Severe intoxication	Very high concentrations (>5000 mg/L) (110 mmol/L) in a severely obtunded patient may be an indication for haemodialysis
	2. Intoxication (actual or possible) in head-injured patients	Low concentrations (<1500 mg/L) (32 mmol/L) help to exclude ethanol as a cause of behavioural disturbance or altered consciousness level
Ethylene glycol	3. Monitoring of treatment of methanol or ethylene glycol intoxication	Blood ethanol should be maintained at 1000–2000 mg/L (22–43 mmol/L) to inhibit alcohol dehydrogenase
	Suspected toxicity	Confirm the presence of ethylene glycol in blood and monitor clearance with treatment
Iron	Overdose	Serum iron >3 mg/L (55 µmol/L) is associated with significant toxicity. In patients without features of severe poisoning, serial measurements are useful in determining need for treatment
Lead	1. Clinical suspicion of poisoning	May confirm toxicity and indicate need for antidote therapy. Patients with blood concentrations >50 µg/dL (2.4 µmol/L) should be considered for treatment
	2. Monitoring industrial exposure	Employers must reduce lead exposure if an employee's blood concentration is ≥50 µg/dL. If ≥60 µg/dL work may be suspended
Lithium	1. Acute overdose	Plasma concentrations may confirm ingestion and indicate possible toxicity, but clinical manifestations are the main guide to toxicity and need for haemodialysis (patient may be asymptomatic with plasma lithium >5 mmol/L, depending on time since ingestion)
	2. Therapeutic monitoring	Therapeutic range 0.4–1.2 mmol/L (but see text). Toxicity in therapeutic use can occur at 1.5–2 mmol/L (often due to interactions with diuretic or NSAID therapy). Severe toxicity >4 mmol/L
Methanol	Suspected toxicity	Confirm the presence of methanol in blood and monitor clearance with treatment
Opioids	1. Overdose	Tests can confirm exposure; treatment is based on clinical decision
Osmolality (plasma)	2. Abuse	Tests can confirm exposure
	3. Toxicity in renal failure patients	Requires measurement of morphine 6-glucuronide
	Suspected methanol or ethylene glycol ingestion	If methanol/ethylene glycol cannot be directly measured then the plasma osmolar gap can be used as a surrogate marker for the presence of toxic alcohol in blood. As the toxic alcohol is metabolized to acid, the osmolar gap falls and may be normal if the patient presents late after ingestion

(Continued)

TABLE 40.3 Indications for measurement of drugs and poisons or other biochemical tests in actual or suspected poisoning (Continued)

Toxin or measurement	Indications for measurement	Interpretation
Paracetamol	1. Paracetamol overdose	High concentrations are an indication for antidote administration according to nomogram To exclude co-ingestion of paracetamol
Paraquat	2. Comatose patient with suspected drug overdose 1. Suspected ingestion (or other route of exposure)	Plasma paraquat concentrations can confirm ingestion and severity of exposure. Positive urine dithionite test confirms exposure. Dithionite test can be carried out on plasma – if positive, confirms massive ingestion Severe toxicity >50 mg/L (215 µmol/L) Concentrations may be in the range 30–60 mg/L (130–260 µmol/L) with minimal impairment
Phenobarbital	1. Overdose 2. Abuse	Therapeutic range 19–37 mg/L (80–160 µmol/L) Ataxia may occur at plasma concentrations >30 mg/L. Dysarthria and lethargy at >40 mg/L. Plasma concentration needs to be corrected for albumin concentration
Phenytoin	3. Therapeutic monitoring 1. Overdose	
Salicylate (aspirin)	2. Following overdose to decide when therapy should be resumed 3. Therapeutic monitoring 1. Salicylate overdose	Therapeutic range 10–20 mg/L (40–80 µmol/L) Provides an indication of severity and action to be taken. May indicate need for haemodialysis if >900 mg/L (6.4 mmol/L) To exclude co-ingestion of salicylate
Solvents	2. Comatose patient with suspected drug overdose 1. Suspected poisoning in an unconscious patient 2. Behavioural disturbance or suspected abuse	May provide diagnosis May confirm exposure
Theophylline	3. Industrial exposure 1. Acute overdose	Tests can confirm excessive exposure High concentrations of >80 mg/L (440 µmol/L) in a severely symptomatic patient (convulsions, cardiac arrhythmias) may be an indication for haemoperfusion Toxicity may occur at 30 mg/L (170 µmol/L) in regular therapy. (Therapeutic range 10–20 mg/L)
Toxicology screen	2. Therapeutic monitoring	
	1. Suspected acute poisoning in an unconscious patient 2. Behavioural disturbance due to suspected abuse or poisoning 3. Confirmation of brain stem death when drug administration is known or suspected	May provide diagnosis May provide diagnosis Presence of potentially toxic drugs prevents diagnosis of brain stem death

MDA, 3,4-methylenedioxyamphetamine; MDMA, 3,4-methylenedioxymetamphetamine; NSAID, non-steroidal anti-inflammatory drug.

drug therapy. However, if convulsions are repeated or prolonged, diazepam is the first-line drug. Drugs such as phenytoin, clomethiazole and thiopental may be required if there is no response to diazepam. In tricyclic antidepressant poisoning, administration of sodium bicarbonate may relieve convulsions. Phenytoin is contraindicated in tricyclic poisoning because both tricyclics and phenytoin are sodium channel blockers and these increase the risk of cardiac arrhythmias.

Body temperature

The core temperature should be recorded with a low-reading thermometer, since hypothermia may complicate poisoning with sedative and antidepressant drugs. The patient should be wrapped in a space blanket, but active rewarming is not usually required. The prognosis is better than in accidental hypothermia from other causes, and temperatures as low as 22 °C are compatible with full recovery. In hyperthermia (rectal temperature >39 °C),

reduction of body temperature is a priority. Depending on the cause of the problem, different treatments may be required. Hyperthermic patients poisoned with agents that produce severe muscle rigidity may require elective paralysis with a muscle relaxant and mechanical ventilation. Malignant hyperthermia or neuroleptic malignant syndrome can be treated with intravenous dantrolene.

Renal complications

Renal function should be monitored (bladder catheterization may be necessary if there is oliguria or severe hypovolaemia). Acute kidney injury may occur as a result of a direct toxic effect or secondary to acute haemolysis with haemoglobinuria or rhabdomyolysis with myoglobinuria. There is controversy as to whether myoglobin and haemoglobin are directly toxic to the kidneys. Most probably they cause an obstructive lesion; alkalinization of the urine increases myoglobin excretion and may prevent acute kidney injury.

General supportive care

In the severely ill patient, general supportive care may be life saving. Regular turning will be needed to prevent pressure necrosis of tissues, and care of the eyes and mouth will also be required.

Intestinal decontamination

Decontamination of the gut following ingestion of poisons is controversial. Syrup of ipecacuanha to induce emesis is now obsolete; there is little evidence that it is effective in emptying the stomach. In the comatose patient, gastric lavage is still occasionally used. However, this procedure may wash some of the poison further down the gut, and it is now rarely recommended.

The most widely used method of gut decontamination is to administer activated charcoal, which adsorbs almost all drugs and poisons. The main exceptions are iron and lithium, which are poorly adsorbed, and alcohols and glycols, whose molar load usually exceeds the adsorptive

capacity of the charcoal. Repeated doses of activated charcoal may be used for sustained-release preparations that have delayed absorption, and may also be used to eliminate certain poisons from the body if there is significant enterohepatic recycling (e.g. carbamazepine and theophylline).

Antidotes

Antidotes form an important part of the management of certain poisons. In some cases (as with paracetamol, opioids, cardiac glycosides, organophosphates and snake bites), they may be life-saving. However, it should be noted that antidotal treatment is only required in a minority of poisonings. The most debated areas with regard to antidotes concern their effectiveness in 'late' paracetamol poisoning (when the patient presents more than 12 h after ingestion) and their effectiveness and use in the management of poisonings by metals. A list of commonly used antidotes is given in [Table 40.4](#).

TABLE 40.4 Commonly used antidotes

Poison	Antidote	Mechanism of action
Anticoagulants (warfarin type)	Vitamin K (phytomenadione)	Competitive antagonist at site of active prothrombin production in liver
Benzodiazepines	Flumazenil	Competitive antagonist at benzodiazepine receptors (often not recommended as may worsen toxicity by causing convulsions)
β -Blockers	Isoprenaline	Competitive antagonist at β -receptors.
Carbon monoxide	Glucagon Oxygen (normobaric or hyperbaric)	Stimulates myocardial adenylyl cyclase Competitive displacement of carbon monoxide from haemoglobin, myoglobin and cytochrome molecules
Cyanide	Dicobalt edentate Sodium nitrite	Chelates cyanide Forms methaemoglobin, which combines with cyanide
Digoxin, digitoxin	Sodium thiosulfate Hydroxocobalamin	Substrate for enzymatic detoxification of cyanide Combines with cyanide to form cyanocobalamin
Ethylene glycol	Fab antibody fragments Ethanol	Antidote forms an inert complex with drug Competitive substrate for alcohol dehydrogenase, slows toxic metabolite production
Heavy metals (lead, mercury, arsenic)	Fomepizole (4-methyl pyrazole)	Inhibits alcohol dehydrogenase, slows toxic metabolite production
Hydrofluoric acid	2,3-Dimercaptosuccinic acid (DMSA) 2,3-Dimercapto-1-propanesulfonate (DMPS)	Chelating agent Chelating agent
Iron salts	Sodium calcium edentate Dimercaprol	Chelating agent Chelating agent
Methanol	D-Penicillamine Calcium gluconate Desferrioxamine Ethanol	Chelating agent Forms an inert complex (calcium fluoride) Chelating agent Competitive substrate for alcohol dehydrogenase, slows toxic metabolite production
Methaemoglobin	Fomepizole (4-methyl pyrazole)	Inhibits alcohol dehydrogenase, slows toxic metabolite production
Narcotic analgesics (heroin, methadone etc.)	Methylthioninium chloride (methylene blue) Ascorbic acid	Cofactor for reduction of methaemoglobin by NADPH Reducing agent
Organophosphates	Naloxone	Competitive antagonist at opioid receptors
Paracetamol	Atropine Pralidoxime; obidoxime Acetylcysteine	Competitive antagonist at acetylcholine receptors Cholinesterase reactivators Accelerates detoxification of potentially toxic metabolite
Thallium	Berlin (Prussian) blue	Chelating agent

Elimination techniques

In some types of poisoning, active elimination techniques are indicated. This approach is most widely used in poisoning by salicylates, theophylline, ethylene glycol and methanol. The techniques include repeated oral doses of activated charcoal, enhanced renal elimination (by alkalization of the urine or diuresis), haemoperfusion and haemodialysis. Haemodialysis may be required for the elimination of the poison (most notably in the case of salicylates, ethylene glycol and methanol) or for the support of renal function. Exchange transfusion may be used in infants. Peritoneal dialysis, plasmapheresis and continuous arteriovenous haemofiltration are less useful in removing poisons, although the latter is an effective technique to support renal function.

SPECIFIC POISONS

Paracetamol (acetaminophen)

Although paracetamol is safe when taken in the recommended dose, it is potentially very toxic in overdose and, since it is widely available over the counter, it is currently the commonest cause of admissions to hospital for poisoning and also of acute hepatic necrosis. It causes approximately 150 deaths each year in Britain.

Mechanisms

Paracetamol is rapidly absorbed from the upper gastrointestinal tract and the majority is metabolized by conjugation with sulfate or glucuronate to non-toxic derivatives. However, a small proportion (approximately 8–10%) is metabolized by a specific cytochrome P450 enzyme, CYP2E1, to produce a highly reactive intermediate, N-acetyl-*p*-benzoquinoneimine (NAPQI). This reactive metabolite can be metabolized by conjugation to form non-toxic mercapturic acid conjugates, provided glutathione is present in the liver cell. When an overdose of paracetamol is taken, the rate of production of NAPQI may exhaust existing glutathione stores and the capacity of the liver to synthesize glutathione (Fig. 40.2). In this case, NAPQI covalently binds to sulfhydryl groups in hepatocytes, forming an irreversible complex, which can result in acute centrilobular necrosis of the liver. Since paracetamol is also metabolized in the cells of the renal tubules, a similar process can also occur in the kidneys, leading to acute kidney injury. There is usually a small amount of renal damage in the presence of hepatic damage, but occasionally renal damage predominates and acute kidney injury may rarely be the presenting feature of paracetamol poisoning.

Toxic dose

The upper therapeutic limit for an adult of 4g daily in divided doses is generally accepted as being safe, but doses above this can cause severe hepatotoxicity and possibly death. However, owing to the wide variation in the metabolic handling of paracetamol by the body, much larger overdoses (>50g) may have little effect in some

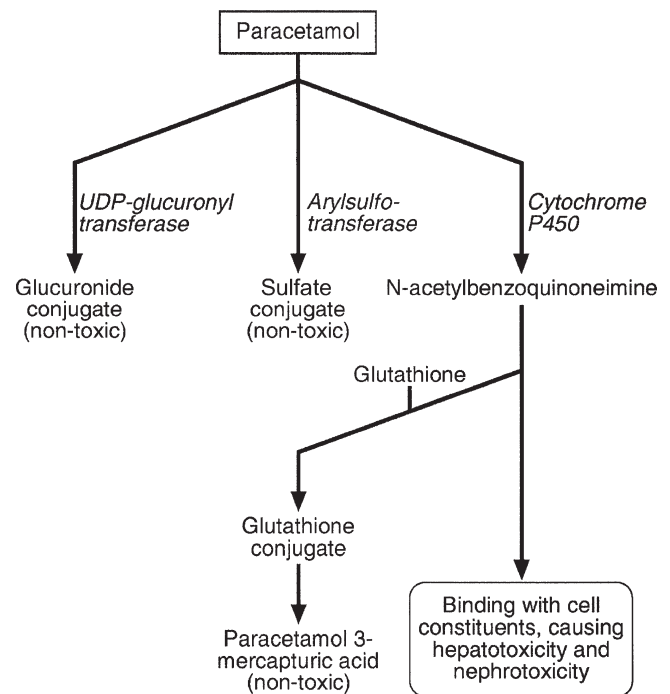


FIGURE 40.2 ■ Metabolism of paracetamol. Induction of the cytochrome P450 pathway increases the production of the toxic metabolite, N-acetyl-*p*-benzoquinoneimine, causing depletion of glutathione stores; a protein-depleted diet leading to deficiency of amino acids may make less glutathione available. In both cases, increased toxicity may result following an overdose.

individuals, producing nothing more than a minor rise in plasma aminotransferase activities.

Clinical features

In the first hours after ingestion of an overdose of paracetamol, symptoms may be minimal, unless it has been taken with some other drug, for example in a compound formulation with dihydrocodeine. There may be malaise, nausea and vomiting. A very large overdose may cause depression of consciousness or metabolic acidosis. By 24–36h, there may be pain in the right hypochondrium. Plasma aminotransferase activities will rise at this time, but the peak values are a poor indicator of prognosis. The prothrombin time (or international normalized ration, INR) is a reliable prognostic indicator of severe hepatotoxicity. Hypoglycaemia and coagulation defects may complicate the hepatic failure. By about 48h, the early signs of hepatic encephalopathy may appear. Death usually occurs after five or six days.

Management

Giving an antidote is more important than emptying the stomach or giving activated charcoal. After a single acute overdose, an antidote should be given if the patient's plasma paracetamol concentration is on or above the line on the treatment nomogram (Fig. 40.3). The need for treatment following repeated suprathreshold dosing (staggered overdose) cannot be assessed using the nomogram and treatment decisions are based on the dose ingested and the patient's body weight. Repeated plasma paracetamol

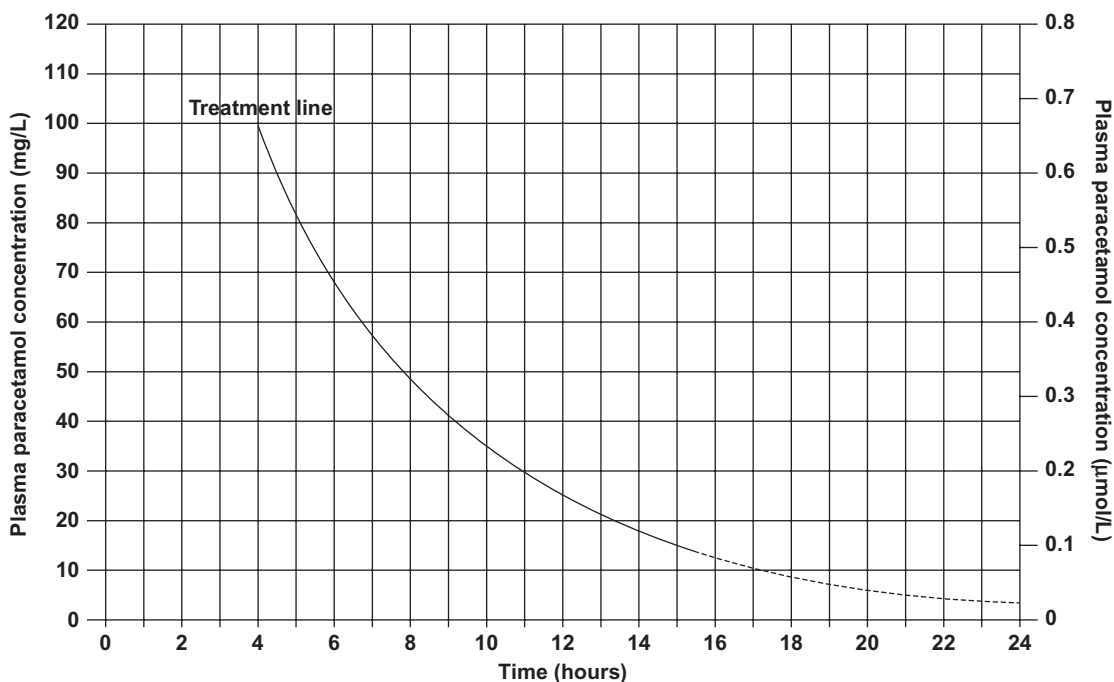


FIGURE 40.3 ■ UK nomogram for treatment of paracetamol poisoning.

estimations should not be performed routinely, but may be helpful in doubtful cases.

Acetylcysteine is the antidote of choice and the intravenous route is the only reliable method of treatment in a patient who is comatose or vomiting. Although the effectiveness of acetylcysteine diminishes considerably with time once 12 h have passed since ingestion, later administration improves outcome even in patients with established liver failure. Acetylcysteine should also be given if the time of the overdose is in doubt or not known. Its main adverse effect is anaphylactoid reactions, which respond to administration of antihistamines and discontinuation or slowing of the infusion.

Salicylate

Although the use of aspirin (acetylsalicylic acid) as an analgesic is decreasing, the management of salicylate poisoning remains a major challenge. The potentially lethal dose in adults is between 24 and 30 g, but death can occur in children under 18 months, from as little as 300 mg. Most fatal poisonings occur in elderly people, because they lack the metabolic reserves to cope with salicylate poisoning and also because attempted suicide with aspirin is commoner in old age. The use of aspirin preparations in children under 16 years of age is largely restricted to rheumatological indications because of the risk of Reye syndrome in children given aspirin for viral illnesses. However, this risk is very low.

Mechanisms

The mechanisms involved in salicylate poisoning are multiple and complex. The clinical features are mainly due to gastrointestinal irritation, stimulation of the respiratory

centre causing respiratory alkalosis and uncoupling of oxidative phosphorylation, leading to heat production and metabolic acidosis. Many other mechanisms are also involved in salicylate poisoning. These are summarized and correlated with the clinical features in Figure 40.4.

Clinical features

Salicylate poisoning usually presents with nausea and vomiting, increased rate and depth of respiration, sweating, tinnitus and sometimes deafness. Consciousness is preserved initially, but confusion, disorientation and loss of consciousness may occur, and usually indicate severe poisoning with a poor prognosis. The patient is often severely volume depleted because of vomiting, hyperventilation and sweating. Excess heat production usually causes compensatory sweating in adults but there may be hyperpyrexia in children. There is usually a combined compensated metabolic acidosis and respiratory alkalosis, with an arterial $[H^+]$ of 32–40 nmol/L (pH 7.4–7.5). Later, the $[H^+]$ rises (pH falls) as the alkali reserve diminishes, and this is a serious sign. In infants, metabolic acidosis may predominate from the start.

Salicylate overdose increases pulmonary capillary permeability, which, in severe poisoning, may cause non-cardiogenic pulmonary oedema. This presents initially as lowering of the arterial oxygen tension and subsequently becomes apparent on chest radiographs. Acute kidney injury is rare and is more likely to occur in children than in adults.

Laboratory measurements

Plasma salicylate concentration should be measured on presentation and at intervals of 4–6 h thereafter, until it has fallen below the toxic range. This is especially

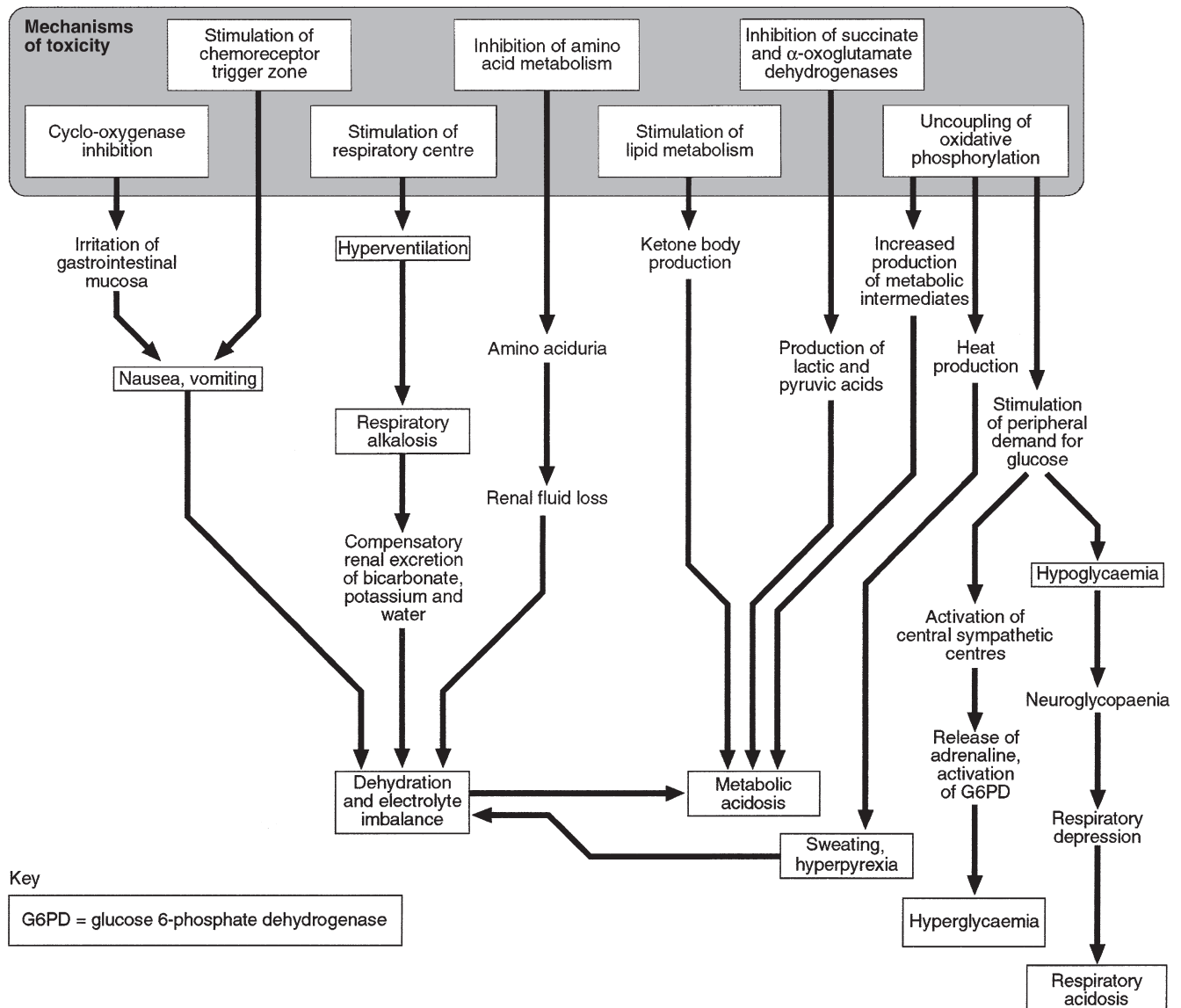


FIGURE 40.4 ■ Pharmacological and metabolic effects of salicylates, leading to biochemical disturbances and clinical effects.

important, since salicylates are precipitated in an acidic environment and may therefore be deposited in the stomach, resulting in delayed absorption. Plasma concentrations may continue to rise for many hours, particularly in serious overdoses. Experience has shown that in most cases with a fatal outcome, plasma salicylate concentrations have risen progressively following admission to hospital.

Management

Activated charcoal should be given to patients presenting to hospital soon after overdose (within 1 h) and a second dose may be given if plasma salicylate concentration continues to rise. Gastric lavage is rarely, if ever, indicated but may be considered if a patient has presented early to hospital after taking a massive overdose. All symptomatic patients should have central venous pressure monitored, and fluid replacement and correction

of electrolyte imbalances, especially hypokalaemia, are priorities. Hypoglycaemia may need to be corrected. Elimination of salicylate can be promoted by administration of sodium bicarbonate to produce a urine pH of >7.5 ($[H^+] < 32$ nmol/L). This has been clearly demonstrated, but the concept of 'forced diuresis' is now no longer used, since the excretion of salicylate is not influenced by urine volume. A urinary alkalization regimen is recommended if the plasma salicylate concentration is greater than 500 mg/L (3.6 mmol/L). Haemodialysis may be necessary in severe poisoning with a plasma salicylate concentration >900 mg/L (6.4 mmol/L), or lower than this if complications are present. As well as removing salicylate, haemodialysis can also correct acid-base and electrolyte imbalances and is therefore preferred to charcoal haemoperfusion. The prothrombin time should be measured; it is rarely prolonged, but if it is, phytomenadione (vitamin K) should be administered intravenously.

Chloroquine

In overdose, this drug can produce hypokalaemia ($[K^+]$ often <2 mmol/L), coma, convulsions and sudden cardiovascular collapse. An acute overdose of >5 g, systolic blood pressure <85 mmHg, a QRS complex duration >0.12 s and a blood chloroquine concentration >25 μ mol/L are all regarded as predictors of a fatal outcome. Active removal methods such as haemoperfusion and haemodialysis are ineffective. Aggressive resuscitation with the use of intravenous noradrenaline (norepinephrine) and diazepam may be life-saving. After treatment, severe hyperkalaemia may ensue, particularly if potassium salts have been given to correct hypokalaemia.

Digoxin

Clinical features

Digoxin and other cardiac glycosides, including plant toxins from oleander (*Nerium oleander*) and yellow oleander (*Thevetia peruviana*), increase the force of contraction of the myocardium. Overdose increases the irritability of ventricular muscle, resulting in extrasystoles, ventricular tachycardia and fibrillation. Conduction is depressed and sinus bradycardia and various degrees of block may also occur. Since digoxin inhibits Na^+, K^+ -ATPase, hyperkalaemia is a feature of digoxin poisoning. A plasma potassium concentration >5.3 mmol/L (in the absence of any other cause of hyperkalaemia) suggests severe poisoning with an increased likelihood of cardiac toxicity. Occasionally, the plasma potassium concentration itself may rise rapidly to life-threatening levels and require urgent treatment.

Other clinical features of acute poisoning include headache, nausea, vomiting, abdominal discomfort, confusion, disorientation and visual blurring or distorted colour vision.

Management

Activated charcoal effectively adsorbs digoxin and should be used if the patient is not vomiting. Atropine may reverse bradycardia. Marked hyperkalaemia should be reversed with intravenous insulin and glucose. There is a concern that administration of intravenous calcium to counteract the effects of hyperkalaemia could worsen digoxin toxicity: however, observational studies do not demonstrate any adverse effects. Whenever there is evidence of severe poisoning such as cardiovascular collapse or refractory hyperkalaemia, digoxin-specific Fab antibody fragments should be obtained urgently. Given intravenously in an appropriate dose, this treatment should produce reversal of signs of toxicity within 20–30 min. These Fab antibodies interfere with commonly available assays for digoxin, so measurements of serum digoxin concentration after treatment will not be reliable.

Iron

Toxicity

Iron absorption is normally regulated by the intestinal mucosa. In overdose, iron-containing medication is corrosive and damages the mucosa, so that iron ions are

absorbed in toxic amounts. When the plasma iron exceeds the iron binding capacity of transferrin, unbound iron circulates freely and may damage the liver, kidneys, cardiovascular system and central nervous system, leading to multi-organ failure.

Toxic doses in terms of elemental iron content are:

- mild-to-moderate poisoning: >20 mg/kg
- severe poisoning: >75 mg/kg
- lethal dose: >150 mg/kg.

Clinical features

The clinical course of iron poisoning can conveniently be divided into four phases. During the first 6 h there may be vomiting, abdominal pain and diarrhoea, resulting from direct irritation of the gastrointestinal mucosa. The vomitus or stools may be dark or blood stained and may smell metallic. Severe fluid loss can lead to lethargy, convulsions, coma, metabolic acidosis and shock. Leukocytosis ($>15 \times 10^9/L$) and hyperglycaemia (>8.3 mmol/L) indicate severe poisoning. This phase usually resolves after 6 h and the patient is often relatively well for the next 18 h. By about 24 h, the patient's condition may worsen: this indicates severe poisoning. Features in this third phase include lethargy, coma, convulsions, cardiovascular collapse, metabolic acidosis, hypoglycaemia and renal and hepatic failure with coagulation abnormalities. Finally, a fourth phase of poisoning may follow the acute episode; it consists of stricture formation that may lead to small bowel obstruction 2–5 weeks after poisoning.

Analysis

Blood should be taken 2–6 h after ingestion. Measurement of serum iron concentration after 6 h may underestimate the amount of free iron because of distribution into tissues. Serum iron concentrations in excess of total iron binding capacity indicate severe poisoning, but the estimation of iron binding capacity may be unreliable in overdose because of the shortcomings of the methods used. The serum iron concentration at 4 h after ingestion is the most reliable guide:

- <3 mg/L (55 μ mol/L): mild toxicity
- 3 – 5 mg/L (55 – 90 μ mol/L): moderate toxicity
- >5 mg/L (>90 μ mol/L): severe toxicity.

Treatment

Chelation therapy. Desferrioxamine binds free iron and the non-toxic complex is excreted by the kidneys. The decision to use it should be based upon the patient's clinical condition and on laboratory analysis. Once parenteral desferrioxamine has been given, colorimetric assay methods for iron are misleading because they measure both free and chelated iron.

Other metals

Poisoning with lead, mercury, bismuth, thallium or arsenic is uncommon, but may arise from deliberate, accidental or occupational exposure. Lead poisoning in children

may be due to pica or follow the use of surma eye make-up in Asian children; in adults it is usually the result of occupational exposure.

The characteristic features of lead toxicity are anaemia with punctate basophilia of the red cells, peripheral muscle weakness, abdominal pain, colic and constipation, and encephalopathy if concentrations are high. Chronic poisoning may produce learning difficulties in children. Chelating agents such as 2,3-dimercaptosuccinic acid (DMSA) and 2,3-dimercaptopropane sulfonate (DMPS) are more effective than the traditional sodium calcium edetate and other chelating agents. DMSA and DMPS are water-soluble analogues of dimercaprol. DMSA is more effective in chelating lead, while DMPS can be used for mercury. Important advantages of DMSA and DMPS over the older chelating agents are that they can be given orally as an outpatient treatment and do not interfere with zinc and copper metabolism.

Orthopaedic implants, such as hip replacements, may contain cobalt and chromium metal. If the implant is functioning well, blood metal concentrations will be very low. However, elevated concentrations >119 nmol/L (cobalt) or >134.5 nmol/L (chromium) are associated with a failing implant and may lead to local tissue reactions, with malignant tumours being described in case reports. In the UK, the Medicines and Healthcare Products Regulatory Agency has issued guidance about the follow-up of patients who have specific types of hip implants. They recommend at least annual assessment of patients during the life of the implant, with more frequent assessment if whole blood concentrations of either chromium or cobalt are above the defined limits. It is unclear if elevated blood cobalt or chromium concentrations produce systemic toxicity in the setting of a failing hip replacement. However, hip revision should be considered if they are persistently elevated.

Organophosphates

Organophosphate poisoning is a major public health issue worldwide resulting in 200 000 deaths each year. Most of these occur through ingestion with suicidal intent. In developed countries, however, the number of poisonings is low and deaths are very rare. The most serious poisoning occurs by ingestion; cutaneous absorption and inhalation of sprays rarely cause serious toxicity.

Toxicity

The acute toxicity of these compounds is due to the inhibition of the enzyme acetylcholinesterase by phosphorylation, resulting in an accumulation of acetylcholine at postganglionic parasympathetic nerve endings (muscarinic receptors), parasympathetic ganglia (nicotinic receptors) and neuromuscular junctions (nicotinic receptors). All the organophosphates inhibit both red cell acetylcholinesterase and plasma cholinesterase (pseudo-cholinesterase) and this provides the basis for biological monitoring of toxicity (red cell measurements being preferred for acute toxicity). For monitoring of occupational exposure, regular measurement of plasma cholinesterase

activity should be carried out: a reduction of the pre-employment value by 30% indicates excessive exposure and the worker should be removed from exposure pending recovery of enzyme activity.

Phosphorylated acetylcholinesterase is relatively stable, which means that the features of poisoning may persist longer than the presence of the organophosphate in the bloodstream. Spontaneous reactivation of the enzyme depends mainly on the chemical structure of the organophosphate. Exposure to compounds such as dimethylphosphates and dimethylphosphorothioates leads to dealkylation of the enzyme (a process referred to as 'ageing'), which makes the enzyme inaccessible to reactivation, either spontaneously or by administration of reactivating agents such as pralidoxime. In poisoning by carbamates, the affected acetylcholinesterase undergoes rapid spontaneous reactivation.

Clinical features and management

Early symptoms of acute exposure to organophosphates are non-specific but lead on to more characteristic features. In mild-to-moderate poisoning there may be headache, blurred vision, miosis, excessive salivation, lacrimation, sweating, wheezing and lethargy. The patient should be kept under observation for at least 24h. Severe poisoning may cause coma, convulsions, respiratory muscle paralysis, bradycardia and hypotension.

The first step is to maintain a clear airway and ensure adequate ventilation, after which atropine should be given until atropinization is achieved, that is, the heart rate is >80 /min, secretions are inhibited or pupils are enlarged.

Pralidoxime (a specific cholinesterase reactivator) should ideally be started within 4h of exposure. In patients who present late to hospital, enzyme inactivation becomes less reversible and pralidoxime is unlikely to have any effect if given after 24–36h.

The patient may relapse after apparent recovery, as a result of an acute myopathy that is distinct from the acute toxicity or the delayed neuropathy that may occur after several weeks. This has been called the 'intermediate syndrome'.

Alcohols and glycols

One of the important features about alcohols from the clinical chemistry point of view is that they all cause a raised osmolar gap; comparison of measured and calculated osmolality can be used for the early confirmation of the potential severity of poisoning by alcohols and glycols.

Ethanol (ethyl alcohol)

Ethanol is probably the best known of all toxins. There are 100 deaths per year in England and Wales from acute alcohol toxicity, but the numbers of deaths from other causes, including road accidents and accidents in the workplace and also from chronic hepatic damage and cancers, is considerably greater than this, with the total toll being over 25 000 deaths each year. When alcohol is taken acutely, effects on behaviour can be demonstrated

at blood concentrations as low as 200 mg/L (4.3 mmol/L). Intellectual performance, judgement and coordination are progressively impaired at increasing blood concentrations. The legal limit in the UK for being in charge of a motor vehicle is 800 mg/L (17.4 mmol/L). By 1600 mg/L (35 mmol/L), most people have marked impairment of coordination and are obviously drunk. However, tolerant individuals may have a blood alcohol concentration of >5000 mg/L (110 mmol/L), without apparent behavioural impairment. By contrast, death from acute toxicity may occur in non-tolerant individuals with blood ethanol concentrations in the range 2000–3000 mg/L (43–65 mmol/L). A fatal outcome from acute alcohol toxicity may result from cardiac or respiratory depression or from aspiration of vomit.

The effects of ethanol are increased by sedative and hypnotic agents, and deep coma may occur in a patient who has taken a combination of a relatively small amount of alcohol plus a small overdose of a benzodiazepine. The clinical assessment of a patient with a severe head injury may be confused by alcohol intoxication. In children, convulsions and hypoglycaemia may result from acute alcohol intoxication. Alcohol potentiates rebound hypoglycaemia and may also cause marked hypoglycaemia in adults. Alcoholic ketoacidosis may follow bouts of heavy drinking (see Chapter 5).

Dilute ethanol enhances gastric emptying, but high concentrations (e.g. at the strength found in spirits) may cause gastric irritation and can delay stomach emptying. In rare cases, haemodialysis may be indicated to remove ethanol from the bloodstream, but supportive care is usually sufficient.

Heavy ingestion of alcohol may lead to the clinical picture of alcoholic hepatitis, with fever, leukocytosis and hepatomegaly. More chronic ingestion can lead to hypertriglyceridaemia; hyperuricaemia is also common in chronic heavy alcohol drinkers. A pseudo-Cushing syndrome occurs in a proportion of alcoholics and may be accompanied by an alcoholic myopathy. The commonest pathological sequel of long-term alcohol abuse is alcoholic cirrhosis, which is accompanied by typical physical signs and the biochemical changes of hepatic impairment. Acute bleeding may occur from ruptured oesophageal varices. The most important tests for detecting chronic abuse, apart from the detection of alcohol in a morning blood or urine sample, are plasma γ -glutamyltransferase activity and carbohydrate deficient transferrin concentration (see Chapter 13).

The alcohol-dependent patient who discontinues drinking because of an accident or a hospital admission may develop the alcohol withdrawal syndrome, in which there is a risk of convulsions and death. Initial irritability may develop into frank hallucinations and infusion of a sedative drug such as clomethiazole may be needed.

Methanol (methyl alcohol)

Methanol is a common constituent of car engine antifreezes. Model aircraft fuel also contains methanol, as do some screenwashes, varnishes and thinners. As little as 10 mL of pure methanol in a child and 50 mL in an adult can be fatal, causing profound metabolic acidosis, coma, convulsions and blindness.

Plasma osmolality, acid–base status and renal function should be checked in all patients with methanol ingestion. Direct measurement of blood methanol concentration is preferable to measurement of plasma osmolality (which is a surrogate indicator), but may not be available in a clinically useful timeframe. If plasma osmolality is normal and there is no acidosis (normal osmolal and anion gap), treatment is not required. Enzymatic metabolism of methanol to toxic metabolites can be prevented by either fomepizole (4-methyl pyrazole) or ethyl alcohol administration. Fluid administration may be required to maintain an adequate urine output for elimination of methanol. Acidosis should be corrected with intravenous sodium bicarbonate. Haemodialysis is indicated if visual disturbance is present, there are features of CNS toxicity or severe metabolic acidosis or if the patient develops acute kidney injury.

Ethylene glycol

Ethylene glycol is a common constituent of automobile antifreezes. If ingested, it tends initially to cause signs of alcoholic intoxication followed later by tachycardia, pulmonary oedema, convulsions and then acute kidney injury (see Fig. 40.5). Metabolic acidosis, leukocytosis, hypocalcaemia and crystalluria are characteristic, owing mainly to the metabolic production of oxalate, which then combines with calcium. Oxalate deposition may lead to meningism and acute tubular injury. The presence of calcium oxalate crystals in the urine is typical and is a useful diagnostic feature.

The main principles of treatment are as for methanol poisoning: to delay the metabolism of ethylene glycol to toxic metabolites by administering fomepizole or ethyl alcohol, to correct the metabolic acidosis and to hasten the elimination of ethylene glycol by increasing urine output or, in severe cases, by using haemodialysis. In addition, plasma calcium concentration should be monitored (hypocalcaemia should be corrected if it occurs), as should renal function (acute kidney injury may require haemodialysis).

Drug and substance abuse

Many drugs and chemicals are widely abused and some cause serious medical complications. There are approximately 600 deaths from opioid abuse, 150 from cocaine and 30 deaths from volatile substance abuse each year in the UK and, although other abused substances cause fewer deaths, they remain an important cause of morbidity.

Apart from opioid toxicity, the management of poisoning with this group of agents is mainly symptomatic. Confirmation of exposure is often necessary for diagnostic or medicolegal reasons and Table 40.5 gives a guide to the typical times during which urine tests remain positive. Immunologically based point-of-care urine tests are commonly used for diagnosis of poisoning by drugs such as opioids, cocaine, cannabis, amphetamine and metamphetamine (usually includes 3,4-methylenedioxymetamphetamine, MDMA, see below). Tests are also available for benzodiazepines,

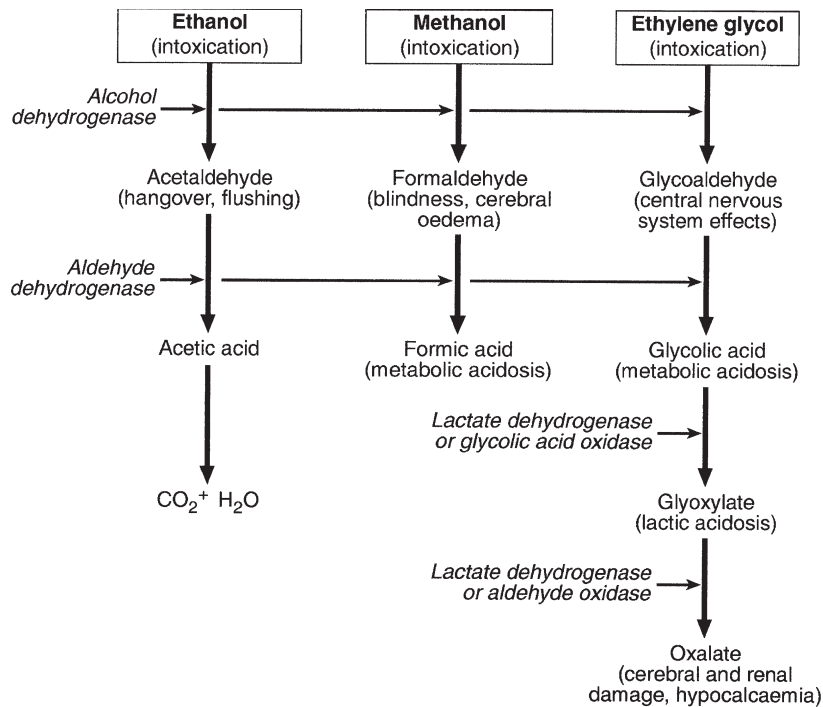


FIGURE 40.5 ■ Metabolism and clinical effects of ethanol, methanol and ethylene glycol. Clinical effects are given in brackets; *enzymes* are in italics. The higher affinity of ethanol for alcohol dehydrogenase is the rationale for its use as an antidote in poisoning with other alcohols. Fomepizole also inhibits metabolism of alcohols by its high affinity for alcohol dehydrogenase.

TABLE 40.5 Duration of positive results in urine after a typical dose taken by a drug abuser

Substance	Limit of detection	Time of detection
Amphetamine	0.25 mg/L	1–2 days
Barbiturates	0.5 mg/L	1–3 days (longer for phenobarbital)
Cannabis	50–300 µg/L	2 days to 3 weeks (depends on usage and limit set)
Cocaine	0.3 mg/L	2–3 days
Codeine	0.25 mg/L	1–2 days
Dihydrocodeine	0.25 mg/L	1–2 days
LSD	0.5 mg/L	2–3 days
MDMA ('Ecstasy')	0.1 mg/L	1–2 days
Methadone	0.25 mg/L	2–5 days
Morphine	0.5 mg/L	1–2 days
Propoxyphene	0.3 mg/L	2–4 days

LSD, lysergic acid diethylamide; MDMA, 3,4-methylenedioxyamfetamine.

methadone (which is structurally different to other opioids and does not give a positive result in tests for opioids), and phencyclidine (ketamine will cross-react in this test).

Amfetamines

The amfetamines are frequently abused for their stimulant effects. Amfetamine sulphate may be injected, inhaled or taken orally. Methylamfetamine ('crystal meth'

or 'ice') comes in the form of crystals that can be smoked and is about 20 times more potent than amfetamine sulphate, but has similar effects.

The main complications of amfetamine toxicity include agitation, convulsions, cardiac arrhythmias, hyperthermia and myoglobinuric acute kidney injury, which can be aggravated by volume depletion. There is no specific pharmacological antidote and management is generally supportive. Sedation and anticonvulsants will be needed for the patient who is agitated or fitting, and maintaining an acidic urine markedly increases the excretion of amfetamine and shortens its elimination half-life. However, if there is a suspicion of rhabdomyolysis, for example following repeated seizures, or if the plasma creatine kinase activity is raised or myoglobinuria is present, then alkalinization of the urine is recommended to prevent myoglobinuric kidney injury.

3,4-Methylenedioxyamfetamine (MDMA)

3,4-Methylenedioxyamfetamine, also known as 'Ecstasy' or 'E', is an amfetamine derivative with different properties from amfetamine; it typically produces euphoria, empathy and a subjective sense of increased energy, which is the reason why it has become popular as a 'dance drug'. First synthesized in 1914, it was briefly used as a mood-modifying agent, but was banned in the UK in 1977. In the usual doses (tablets or capsules contain 30–150 mg; users take between half and five orally during the course of an evening), it has few adverse effects in the majority of people. The commonest effects are

trismus, tachycardia, sweating and agitation. A small proportion of users develop muscle pain and stiffness, which may persist for a week. When the drug is taken before or during strenuous exercise such as dancing, it may lead, in rare cases, to collapse, convulsions and acute hyperthermia, rapidly followed by disseminated intravascular coagulation and rhabdomyolysis. Deaths have occurred after a single dose.

The reason for collapse is related to the combination of physical hyperactivity and inadequate fluid replacement, but it is not known why only a very small number of individuals are affected. Management initially should be directed to restoring fluid volume and reducing body temperature. Intravenous dantrolene may be effective. Ingestion of MDMA may also produce a serotonin syndrome, which is characterized by a triad of altered mental status (agitation, confusion), neuromuscular hyperactivity (clonus) and autonomic instability (hyperthermia, tachycardia). In order to prevent the hyperthermia becoming fatal, paralysis and ventilation may be required. Another complication of MDMA use is hyponatraemia, which usually occurs when excess fluid has been taken but there has been insufficient exercise to sweat off the fluid. This occurs because MDMA causes serotonin production, which in turn causes release of excess arginine vasopressin (antidiuretic hormone) from the posterior pituitary. The patient is likely to be confused and may have convulsions, but has a normal body temperature. Plasma sodium will be ≤ 125 mmol/L. The most important step in management is to avoid giving any fluids; most patients will recover spontaneously.

Although MDMA is taken because it causes the release in the nervous system of large amounts of serotonin, which is one of the neurotransmitters responsible for the 'high' caused by the drug, this leads to depletion of serotonin and a 'midweek low' is a well-known consequence of MDMA consumption. Serotonin neurotoxicity has been demonstrated in animals, and a reduction in cognitive abilities, particularly memory, has been demonstrated in humans following chronic intermittent use.

Heroin (diamorphine)

Heroin abuse is common in the UK, causing over 500 deaths per year. Many of these are from respiratory failure, which may occur after the first use of the drug. Experienced users may die from acute respiratory depression or aspiration of vomit following loss of tolerance after a period of abstinence. Tolerance is rapidly lost; the dose requirement may fall by 10-fold over a period of 2–3 days of abstinence, so that the dose which previously produced euphoria may prove fatal.

The classic signs of opioid toxicity consist of depressed consciousness, which may amount to profound coma, small or 'pinpoint' pupils (pethidine is an exception to this and the atropine in Lomotil[®], a mixture of diphenoxylate and atropine, may cause dilated pupils) and respiratory depression. The pattern of opioid induced respiratory depression is characteristic and consists of a slowing of respiration, whereas other respiratory depressants tend to decrease the depth of respiration with less effect on the rate. An adequate dose of naloxone

should reverse these symptoms; an initial dose may only produce a minimal or partial response and further doses may be required. The presence of other central nervous depressant drugs or other causes of coma may mask the response to naloxone.

Once the diagnosis of opioid toxicity has been confirmed by administration of naloxone, the choice can be made as to whether to continue reversal with an infusion of naloxone sufficient to maintain adequate spontaneous respiration or whether to ventilate the patient until the drug is metabolized. Diamorphine has a short elimination half-life of 2–4h, so that treatment may not need to be prolonged. Other opioid drugs such as buprenorphine, dihydrocodeine, methadone and dextropropoxyphene have considerably longer elimination half-lives and a longer period of recovery should be anticipated.

Complications of heroin overdose include a chemical pneumonitis owing to aspiration of vomit, non-cardiogenic pulmonary oedema and non-traumatic rhabdomyolysis. The question of provoking heroin withdrawal symptoms in an addict by administration of naloxone is often raised; opioid withdrawal is a non-fatal condition, which is, in any case, short-lived because of the short duration of effect of a single dose of naloxone (10–30 min).

Lysergic acid diethylamide (LSD)

Lysergic acid diethylamide is a synthetic hallucinogen, which may lead to a patient being admitted to an intensive therapy unit because of behavioural disturbances, accidents or suicide attempts. There is no specific antidote.

Cocaine

The hydrochloride is typically snorted through a straw or a rolled up bank note, while the free base form, in small crystals or 'rocks', is drawn directly into the lungs by flaming a rock with a cigarette lighter. Cocaine produces marked vasospasm and hypertension, and patients most frequently present to emergency departments with chest pain. They may also present with confusion, convulsions, myocardial infarction, acute heart failure or cerebral haemorrhage. There is no specific antidote, but intravenous diazepam relieves the chest pain and reduces the risk of convulsions. After this, blood pressure may be controlled with agents such as intravenous nitrates. Other management is supportive.

Cannabis

Tetrahydrocannabinol, the active ingredient of the cannabis plant, combines with receptors in the brain and periphery to cause mental effects together with vasodilatation and tachycardia. Effects last a few hours, but the drug can persist in tissues for a long time. Acute overdose is unlikely to be fatal, but heavy use is cumulative and can lead to agitation, hallucinations and paranoid behaviour.

The main psychoactive component of cannabis, Δ^9 -tetrahydrocannabinol (THC), reaches peak concentration in the blood within minutes of smoking, but rapid absorption by fatty tissues then leads to a rapid fall. However,

one of the main metabolites of cannabis, 11-nor- Δ^9 -tetrahydrocannabinol carboxylic acid (carboxy-THC), which is detectable in the blood shortly after use, persists in the blood for a week or more. Immunoassays do not distinguish between active THC and its inactive metabolites. For this reason, blood or urine tests may remain positive for a long time after cannabis consumption, and cannot be taken to indicate intoxication.

Solvents

Volatile substance abuse (solvent abuse, 'glue sniffing') is relatively common in teenagers. Most obtain a sensation of intoxication within seconds of inhalation and come to little harm. The commonest substance abused is toluene. However, if the dose is too high, coma and convulsions may ensue and inhalation of vomit is a possibility. Since volatile substances (particularly chlorinated hydrocarbons) sensitize the myocardium to catecholamines, cardiac arrhythmias may occur and cause a fatal outcome. Butane from cigarette lighter refills may be inhaled directly; the gas may cause asphyxia or its cooling effect may cause cardiac arrest through vagal inhibition.

Chronic complications of solvent abuse are unusual, but brain, liver and kidney damage have all been reported. Carcinogenesis has not been documented. Management is generally symptomatic; cardiac arrhythmias may respond to an intravenous dose of a β -blocker.

Benzodiazepines

The benzodiazepines are widely thought to be non-fatal in overdose, but about 200 deaths occur every year in the UK from respiratory depression, aspiration of vomit or hypothermia. Usually such deaths occur when benzodiazepines are combined with other agents such as ethanol. In most patients with an uncomplicated benzodiazepine overdose, the coma amounts to a deep sleep with an adequate gag reflex and preserved tendon reflexes and response to painful stimuli.

Patients rarely need management in an intensive therapy unit, but sometimes the drug has been co-ingested with ethanol, opioids, barbiturates or tricyclic antidepressants. The question may arise as to whether flumazenil should be given as a therapeutic or diagnostic test. Reversal of benzodiazepine toxicity could provoke convulsions due to simultaneously ingested tricyclic antidepressants. In general, it should only be used in carefully considered circumstances.

Theophylline

Clinical features

Most oral preparations of theophylline are sustained-release formulations. The patient may be asymptomatic on presentation. Symptoms usually take several hours to develop and include vomiting, abdominal pain, haematemesis, irritability and hyperventilation. A sinus tachycardia is usual and there may be hypotension, cardiac arrhythmias (most commonly supraventricular

tachycardia), convulsions, hypokalaemia and hyperglycaemia. Coma is not common. Rhabdomyolysis and acute kidney injury are rare complications.

Management

Activated charcoal in repeated doses is the treatment of choice for severe poisoning. This is indicated if the patient is clinically severely poisoned and/or has a plasma theophylline concentration $>40\text{mg/L}$. In life-threatening poisoning, charcoal haemoperfusion can be used, if available.

Antidepressants

Tricyclic antidepressants

Acute overdose with tricyclic antidepressants is the commonest cause of poisoning admissions to intensive therapy units in the UK. The mechanisms of toxicity are complex and are due to at least three pharmacological effects:

- an anticholinergic effect, which causes delayed gastric emptying, sinus tachycardia and mydriasis
- blockade of noradrenaline uptake at adrenergic synapses, which may produce hypotension
- a quinidine-like or membrane-stabilizing effect. In low doses, this last effect is antiarrhythmic, but in overdose, it delays conduction and depresses contractility. The quinidine-like effect slows sodium flux into cells and is the mechanism underlying cardiac toxicity. It may be aggravated by hypoxaemia (Fig. 40.6).

Clinical features. There may be initial nausea, vomiting, agitation and hallucinations, giving way to coma. Tendon reflexes may be equal and very brisk, with extensor plantar reflexes; as coma deepens, muscle tone becomes flaccid and reflexes are lost. The pupils are more often mid-sized than widely dilated. Convulsions are common and respiratory depression may occur.

The most serious and potentially fatal complications are cardiac arrhythmias (most commonly atrioventricular block, ventricular tachycardia or ventricular fibrillation, which may be followed by asystole) and profound hypotension, which may amount to electromechanical dissociation. Pulmonary oedema may occur. The electrocardiogram (ECG) shows a sinus tachycardia and there may be a right bundle branch block pattern. A QRS duration $>100\text{ms}$ is regarded as the best indicator of risk of cardiac toxicity.

Management. Cardiopulmonary status must be assessed urgently and resuscitation commenced if necessary. Poisoning is potentially reversible, provided hypoxic cerebral damage has not occurred. In significant poisoning, the most important treatment is alkalization with intravenous sodium bicarbonate, aiming for a blood $[\text{H}^+]$ of 32nmol/L (pH 7.5). This is indicated even in the absence of acidosis when there is prolongation of the QRS duration on ECG, there are cardiac arrhythmias or resistant hypotension. Seizures should be treated with benzodiazepines. If recurrent, phenobarbital or thiopentone may

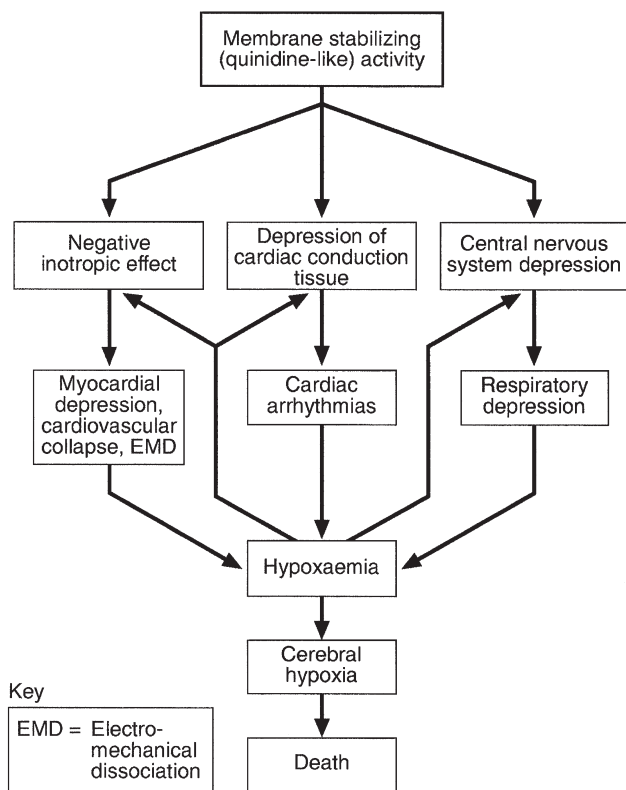


FIGURE 40.6 ■ Mechanisms by which the membrane-stabilizing activity of drugs can cause death.

be needed in a critical care setting. Phenytoin is contraindicated as both tricyclics and phenytoin block cardiac sodium channels. In life-threatening poisoning, lipid emulsion (e.g. intralipid) therapy may be considered; this acts as a lipid reservoir in the circulation that may 'trap' the drug and reduce the concentration at the receptor.

Monoamine oxidase inhibitors

Toxicity. Several drugs and foods are prohibited for patients taking monoamine oxidase inhibitors to prevent the well-known 'cheese' reaction, which consists of a sudden and severe rise in blood pressure owing to tyramine-provoked release of noradrenaline.

Overdose of monoamine oxidase inhibitors produces symptoms that build up over 12–24h, with muscle twitching progressing to widespread muscle spasms, trismus and opisthotonos. The blood pressure may vary between hypotension and moderate hypertension; there is usually a sinus tachycardia and the patient is warm to the touch, sweating profusely and has fixed dilated pupils. The core temperature may rise steeply, leading to death from hyperthermia. The muscle spasms may lead to rhabdomyolysis, which can cause acute kidney injury. Disseminated intravascular coagulation may occur as a complication of hyperthermia.

Management. If the rectal temperature rises above 39°C the patient should be electively paralysed with pancuronium and mechanically ventilated for 12–24h both to reduce heat generation through muscle spasm (and thus correct

hyperthermia) and prevent rhabdomyolysis. Hypotension is usually secondary to hypovolaemia, but dopamine may be tried if fluid replacement fails to restore blood pressure.

Other antidepressants

Lofepramine. This tricyclic drug is metabolized to desipramine, but toxicity in overdose is usually relatively mild and deaths are rare.

Trazodone. This drug is chemically unrelated to the tricyclic antidepressants. Toxicity is usually mild, with drowsiness, dizziness and occasionally coma.

Venlafaxine. This drug is less cardiotoxic than the tricyclic antidepressants, but may cause convulsions in overdose.

Fluvoxamine, fluoxetine, sertraline, paroxetine. These serotonin reuptake inhibitors rarely cause serious toxicity in overdose, though a mild serotonin syndrome may occur, and they can also interact with other drugs to produce a serotonin syndrome. Management is symptomatic.

Citalopram. This serotonin reuptake inhibitor is more cardiotoxic than others, producing QT interval prolongation.

Lithium

Toxicity

Lithium is eliminated by the kidneys and reduced renal function can lead to accumulation. The therapeutic range is narrow (plasma concentration 0.8–1.2mmol/L in the management of hypomania or mania in bipolar affective disorder; 0.4–1.0mmol/L for prophylaxis: note that blood should not be collected into a lithium heparin tube). Toxicity in therapeutic use can be caused by changes in fluid or electrolyte balance (particularly owing to reduction in fluid intake or increased fluid loss from diarrhoea or vomiting). Diuretic therapy or non-steroidal anti-inflammatory drugs may also cause toxicity by reducing lithium excretion.

Clinical features

Symptoms of toxicity include confusion, agitation, drowsiness, tremor, hyper-reflexia, hypertonia, ataxia, vomiting, convulsions and, rarely, electrocardiographic changes, diabetes insipidus and acute kidney injury. In acute poisoning, serum concentrations of >5mmol/L may be associated with minimal symptoms and the elimination half-life is relatively rapid. The development of marked neurological symptoms is an indication for active elimination. Deaths from lithium poisoning are rare, but neurological impairment may be permanent.

Management

Lithium is not adsorbed by activated charcoal but is readily excreted by the kidneys and a high urine output should be ensured. Saline diuresis enhances lithium excretion but urine alkalization does not.

In a symptomatic patient with high serum lithium concentrations after an acute overdose, haemodialysis should be considered. Although efficiently removed by haemodialysis, lithium has a large volume of distribution leading to a 'rebound' rise in plasma concentrations and repeated dialyses may be necessary. Peritoneal dialysis is much less effective.

Cyanide

Cyanide can be rapidly fatal. It may be inhaled as hydrogen cyanide (either pure, as an industrial gas or as a product of combustion in cases of smoke inhalation) or enter the body by ingestion of cyanide salts or intestinal hydrolysis of cyanogenic glycosides (e.g. following massive ingestion of stone fruit kernels or apple pips) or by cutaneous absorption of cyanide salts in industrial situations.

Cyanide acts as a chemical asphyxiant, rapidly blocking cellular oxygen utilization so that cerebral function and circulation are rapidly impaired with the development of a metabolic acidosis. Early signs include hyperventilation and tachycardia, but coma, cyanosis and convulsions soon supervene. Rapid diagnosis and intervention are essential. A high inspired oxygen content (or hyperbaric oxygen, if available) is an effective treatment, though its mode of action is unclear, perhaps being due to simple displacement of cyanide. When the diagnosis is certain, dicobalt edetate can be given intravenously over 1 min, repeated as necessary, and is an effective antidote. Although relatively non-toxic when administered to a patient poisoned with cyanide, this antidote can produce severe anaphylactoid reactions with laryngeal oedema and convulsions if given to a patient who is not poisoned by cyanide. The safest course of action is not to give the antidote unless the patient's level of consciousness is deteriorating. Another group of antidotes (amyl nitrite and sodium nitrite) act by producing methaemoglobin, each molecule of which can bind four cyanide ions. Sodium thiosulphate acts to neutralize cyanide by donating a sulphur group to produce sodium thiocyanate, in a reaction that is catalysed by rhodanese.

Subacute or chronic cyanide toxicity can also occur during prolonged nitroprusside therapy. The main clinical feature is a lactic acidosis – if none is present then cyanide toxicity can be ruled out. Cyanide measurement is not usually necessary: a rapid improvement should occur after administration of sodium thiosulphate intravenously. Other cyanide antidotes are not indicated.

Carbon monoxide

Toxicity

Carbon monoxide is a colourless, odourless gas produced by combustion of carbon-containing compounds. It has a high affinity for haemoglobin and cytochrome enzymes, and is highly toxic because it acts as a chemical asphyxiant. It accounts for around 50 deaths each year in the UK, which may be suicidal from motor vehicle exhaust fumes, or accidental, mainly from release of products of combustion (of gas, solid or liquid fuels) in the home or exposure to fire smoke.

As little as 0.1% carbon monoxide in the inspired air can be fatal over a period of several hours, while 1% can be fatal in minutes. Toxicity is due mainly to combination of carbon monoxide with haemoglobin to form carboxyhaemoglobin, but is also due to interference with the cellular uptake of oxygen as a result of its combining with cytochrome enzymes. Carboxyhaemoglobin in the blood has a half-life of about 4h if the patient is breathing air. The half-life is markedly reduced (to 1h) if the patient is breathing 100% oxygen and is still further reduced (to 20 min) if the patient is breathing oxygen in a hyperbaric chamber at 2.5 atmospheres, but the effectiveness of hyperbaric oxygen in diminishing the toxicity of carbon monoxide is hotly debated.

Clinical features

The symptoms of exposure are non-specific: lethargy, nausea, headache, drowsiness, hyperventilation, leading to vomiting, collapse, coma and convulsions. Elderly people may present with stroke or exacerbation of coronary heart disease precipitated by reduced oxygen-carrying capacity of the blood. Non-traumatic rhabdomyolysis may occur. If the source of exposure is not apparent, the diagnosis may also not be apparent. Once diagnosed or suspected, evidence should be sought, for example a malfunctioning flue from a fire or heating appliance, or motor exhaust fumes. A rare cause is exposure to methylene chloride, which is metabolized to carbon monoxide by hepatic smooth endoplasmic reticulum.

Specific physical signs are minimal; the 'cherry-red' colour of carboxyhaemoglobin is usually only noted at post-mortem. Diagnosis of exposure should be made by measuring carboxyhaemoglobin on a whole blood sample. Correlation of the percentage of carboxyhaemoglobin with clinical effects is poor and is complicated by the disappearance of carbon monoxide from the blood following exposure, while the clinical effects may remain for longer. However, normal values in non-smokers are <1% and in smokers <5%. Carbon monoxide is produced during the catabolism of tetrapyrroles and severe haemolysis in haemolytic anaemia can, rarely, give a carboxyhaemoglobin as high as 8%.

Management is by removal from exposure, resuscitation as necessary and immediate administration of a high inspired oxygen concentration. Blood should be taken at the earliest possible moment for carboxyhaemoglobin estimation. The role of hyperbaric oxygen therapy in carbon monoxide poisoning is controversial but should be considered in life-threatening poisoning, if the treatment is available.

Methaemoglobinaemia

Causes

The more important causes of life-threatening methaemoglobinaemia, likely to present with acute symptoms, are:

- aniline
- sodium or potassium nitrite
- amyl or butyl nitrite
- sodium or potassium nitrate
- sodium or potassium chlorate.

Nitrites are more potent methaemoglobin formers than nitrates, but nitrate can be converted into nitrite by intestinal bacteria. Inhaled nitrites (amyl and butyl nitrites) produce little methaemoglobin unless the liquid is swallowed, when severe toxicity may occur. Many drugs, including sulphonamides, dapsone and local anaesthetics, can increase methaemoglobin production but are unlikely to produce concentrations exceeding 30%. It should also be noted that the cause may be iatrogenic, for example sodium nitrite given for cyanide poisoning.

Symptoms

At low concentrations, the patient may be deeply cyanosed but otherwise asymptomatic and require no treatment. At high concentrations, the patient may be dyspnoeic on exertion and have postural hypotension. At severely toxic concentrations, the patient becomes flaccid and comatose. Cardiac arrhythmias and convulsions may occur and may progress to cardiorespiratory arrest. Methaemoglobinaemia should be suspected if the skin has a blue or greyish, cyanosed appearance: the blood may be a dark or chocolate-brown colour. This colour change occurs with methaemoglobin concentrations of 15–20%, but clinical symptoms only appear at values >20–30%, while consciousness is likely to be lost at >50%. Death is common at >70%.

Management

In addition to supportive measures, which should include 100% oxygen, methaemoglobinaemia may be reversed by the administration of intravenous methylthionium chloride (methylene blue). It should only be given to symptomatic patients. When the cause is chlorate poisoning, methylene blue is unlikely to reverse the methaemoglobinaemia, although ascorbic acid can be given. In extreme cases, exchange transfusion can be used, though the transfused blood may also become affected and prolong the methaemoglobinaemia. Hyperbaric oxygen, if available, may allow sufficient oxygen to be transported in solution in the plasma to maintain life.

Plant and fungal toxins

Most plants are non-toxic and most childhood ingestions of plant material do not lead to toxicity. However, many plants contain chemical agents that are potentially poisonous when accidentally ingested by children or taken as food by adults; in addition, suicidal or homicidal ingestion may occur. Two examples have already been given in this chapter: *Thevetia* and oleander contain cardiac glycosides, and the seeds of a number of varieties contain cyanogenic glycosides, which may cause cyanide poisoning if sufficient seeds or kernels are ingested. Other potentially serious plant poisons are included in Table 40.6.

Acute fungal poisoning most commonly arises from eating toxic species in mistake for edible species. Many mushrooms can cause gastrointestinal symptoms shortly after ingestion, but most of these are not associated with serious poisoning; a rule of thumb is that mushroom poisonings with symptoms starting <6h after ingestion are unlikely to be due to highly toxic species. A few species, especially *Coprinus atramentarius*, can produce an Antabuse®-like reaction if ingested with alcohol, because they contain a chemical that inhibits aldehyde dehydrogenase.

The most important group of poisonous mushrooms includes those which contain cyclopeptides known as amanitins (*Amanita phalloides*, *A. verna*, *A. virosa* and some *Galerina* and *Lepiota* species). After an initial period of severe diarrhoea, beginning 6–12h after ingestion, the patient may develop fulminant hepatic failure: a single mushroom can cause death and the mortality rate from ingestion is 20%. Another important and potentially fatally poisonous mushroom species is *Gyromitra esculenta*, which contains gyromitrin. If this chemical is not previously removed by evaporation or boiling, severe toxicity may result from the metabolic conversion of gyromitrin to monomethylhydrazine, which is a competitive inhibitor of pyridoxal phosphate. Symptoms start suddenly 6–12h after ingestion and consist of headache, vomiting, diarrhoea and convulsions. Liver damage may also occur. The mushroom species *Cortinarius* contain the toxin orellanine that causes acute kidney injury, which may require renal replacement therapy, after an initial latent, symptomless, phase of several days.

TABLE 40.6 Plants that cause potentially serious poisoning

Plant	Toxic principle	Possibly fatal amount (in an adult)
Ackee (unripe fruit)	Hypoglycin	One fruit
Apple or pear seeds (<i>Malus</i> , <i>Pyrus</i> spp.)	Amygdalin (cyanogenic glycoside)	100 seeds
Castor oil plant (<i>Ricinus communis</i>)	Ricin	Two beans
Death cap (<i>Amanita phalloides</i>)	Amanitin, phalloidin (cyclopeptides)	One mushroom
<i>Galerina</i> spp. mushrooms	Amanitin-type cyclopeptide	One mushroom
<i>Gyromitra</i> spp. mushrooms	Monomethylhydrazine	One mushroom
Holly (<i>Ilex aquifolium</i>)	Ilicin	30 berries
Jequirity (<i>Abrus precatorius</i>)	Abrin	One bean
Oleander (<i>Nerium oleander</i>)	Oleandrin (cardiac glycoside)	One leaf
Stone fruit kernels (<i>Prunus</i> spp.)	Amygdalin (cyanogenic glycoside)	30 kernels
Hemlock water dropwort (<i>Oenanthe crocata</i>)	Oenanthe toxin	Several leaves or one root
Yew (<i>Taxus baccata</i>)	Taxine A and B	50 needles

CONCLUSION

This chapter outlines the causes, mechanisms and management of the more important types of human poisoning. Poisons – drugs, chemicals and plant and animal toxins – can disrupt the physiological and biochemical functions of the body in a remarkable variety of ways. Many individual toxins produce a characteristic clinical and biochemical pattern of harm ('toxidrome'), which sometimes provides the key to the diagnosis. Management involves supportive care, decontamination and, in some cases, the use of specific antidotes. The clinical biochemistry laboratory may become involved in the identification and quantification of poisons and also in the monitoring of biochemical responses following poisoning; close communication between clinicians and biochemists is necessary for the efficient diagnosis and management of many types of poisoning.

ACKNOWLEDGEMENT

I would like to acknowledge the contribution of the late John A. Henry, who wrote the chapter for previous editions of this book.

Further reading

Baselt RC. Disposition of toxic drugs and chemicals in man. 9th ed Seal Beach, CA: Biomedical Publications; 2011.

Reference work on toxic and fatal levels of drugs and poisons.

Dart RC. Medical toxicology. 3rd ed. London: Lippincott Williams & Wilkins; 2004.

Large, comprehensive text on human poisoning.

Hathaway GJ, Proctor NH. Proctor and Hughes' chemical hazards of the workplace. 5th ed London: John Wiley; 2004.

Useful and practical book on industrial toxins.

Vale A., Mucklow J., editors. *Medicine* 2012;40:41–104 and 40:105–66.

These two issues of the Medicine journal are devoted to poisoning.

APPENDIX 40.1: POISONS CENTRES

TOXBASE, the primary clinical toxicology database of the National Poisons Information Service in the UK, is available on the internet to registered users at: www.toxbase.org. It provides information on routine diagnosis, treatment and management of patients exposed to drugs, household products and industrial and agricultural chemicals.

Specialist information and advice on the treatment of poisoning is available from the UK National Poisons Information Service by telephone day and night: Tel: 0844 892 0111 (July 2013).

Help with identifying capsules or tablets may be available from a regional medicines information centre.

Metabolic effects of tumours

Wassif S. Wassif • James E. East

CHAPTER OUTLINE

INTRODUCTION 808

NEUROENDOCRINE TUMOURS 808

Carcinoid tumours 808

MULTIPLE ENDOCRINE NEOPLASIA 810

Multiple endocrine neoplasia type 1 810

Multiple endocrine neoplasia
type 2 812

Other familial syndromes associated with
multiple endocrine neoplasia 813

METABOLIC CHANGES IN MALIGNANCY 813

Introduction 813

Paraneoplastic syndromes 814

Cancer cachexia 816

ENDOCRINE SEQUELAE OF TUMOURS AND THEIR TREATMENT 818

Effects on somatic growth 818

Reproductive consequences of therapy 819

CONCLUSION 820

INTRODUCTION

Tumours may exert metabolic effects on the host via a wide range of mechanisms. Some of the metabolic derangements related to tumours are directly related to a hormone or other substance that the tumour secretes. Such hormone secretion can be appropriate to the cell line from which the tumour originates or may be inappropriate or 'ectopic'. In the last decade, molecular genetic techniques have helped characterize tumours and explain how neoplastic tissues can secrete hormones that are not typically associated with them. Metabolic derangements can also arise non-specifically, in the absence of hormone secretion or immunological phenomena, for example owing to tumour burden compromising host metabolism or the effects on the host of rapid cell turnover. Autoimmune reactions evoked by tumour antigens may also give rise to systemic effects and are increasingly being implicated in the causation of various tumour-related or paraneoplastic syndromes. The recognition and characterization of the antibodies to shared tumour and host antigens that play a prominent role in paraneoplastic syndromes have advanced their use in both diagnosis and treatment.

Scientific advances have rapidly opened up new possibilities for the management of tumours, even in their advanced stages, where formerly only observation and palliation was possible. Clinical biochemistry has a key role in both diagnosis and monitoring treatment.

NEUROENDOCRINE TUMOURS

Neuroendocrine cells can be identified by their production of a neurotransmitter, neuromodulator or neuropeptide hormone. Neuroendocrine tumours (NETs) are distinguished by their ability to secrete

peptides causing characteristic endocrine syndromes. The gastroenteropancreatic system has diffuse neuroendocrine components and NETs most commonly arise from these tissues, more than 50% being carcinoid tumours.

Carcinoid tumours

Carcinoids are NETs that arise from enterochromaffin cells, principally in the intestine (~80% in the ileum), the main bronchi and, rarely, other tissues including the ovaries, thymus, pancreas and thyroid. Carcinoid tumours all show a similar pattern of neuroendocrine expression and may contain or secrete amines, peptides or prostaglandins. The carcinoid syndrome is the clinical result of systemic release of these substances, particularly serotonin.

Clinical presentation

The incidence of carcinoid tumours found at post-mortem has been estimated to be as high as 1 in 150, whereas carcinoid syndrome only occurs in approximately 1 in every 50 000 individuals, indicating that most tumours appear to be non-secretory. The products of primary intestinal carcinoids are metabolized in the liver prior to release into the systemic circulation: the development of carcinoid syndrome in patients with a gastrointestinal tumour therefore indicates the presence of hepatic metastases. Carcinoid tumours at bronchial (~10% of the total) and other sites secrete products directly into the systemic circulation and can be associated with carcinoid syndrome without the presence of metastases. The main clinical features of carcinoid syndrome are flushing and diarrhoea (Box 41.1). The diarrhoea tends to be secretory in type and weight loss is common. Patients who flush repeatedly may, in time, develop a cyanotic telangiectatic appearance, which persists. Wheezing may occur, caused by transient

BOX 41.1 Clinical features of carcinoid syndrome

- Flushing 70% (usually without sweating)
- Diarrhoea 50%
- Intermittent abdominal pain 40%
- Valvular heart disease (~20% at presentation)
- Palpitations
- Wheezing
- Facial telangiectasis
- Pellagra-like skin lesions

bronchoconstriction, which is most commonly associated with bronchial carcinoids. Hypertension is not typically a feature of carcinoid syndrome: hypotension is more common. Carcinoid syndrome is also associated with fibrosis; the bowel, retroperitoneum (and hence ureters), lungs and heart can all be affected. Fibrosis is thought not to be linked directly to serotonin but to mitogenic growth factors that drive fibroblast proliferation. Tryptophan is a precursor of both serotonin and nicotinic acid. Thus, in carcinoid syndrome, excess production of serotonin may lead to nicotinamide deficiency, manifesting as pellagra with dermatitis of sun-exposed areas. Carcinoid crisis has been described in certain circumstances including induction of anaesthesia, liver biopsy and following combination chemotherapy. It is manifest as extreme flushing, explosive diarrhoea, labile blood pressure and cardiac rhythm irregularities, including asystole.

Metabolism of serotonin

The rate limiting step in the biosynthesis of serotonin is the hydroxylation, by tryptophan hydroxylase, of tryptophan to 5-hydroxytryptophan (5HTP); this is subsequently decarboxylated to yield 5-hydroxytryptamine (serotonin, 5HT) (see Fig. 41.1). 5-Hydroxytryptamine is stored in neurosecretory granules or secreted into the bloodstream. After secretion, some is taken up into platelets. Oxidative deamination, catalysed by monoamine oxidase and aldehyde dehydrogenase, inactivates 5HT to 5-hydroxyindoleacetic acid (5HIAA), which is excreted in the urine.

The normal excretion of 5-HIAA is <math><50\ \mu\text{mol}/24\text{h}</math>. In patients with typical carcinoids, 99% of metabolized 5HT and 5HTP are excreted as 5HIAA. Urinary 5HIAA excretion is typically in the range 150–1500 $\mu\text{mol}/24\text{h}</math>, generally exceeds 500 $\mu\text{mol}/24\text{h}</math> and occasionally is as high as 3000 $\mu\text{mol}/24\text{h}</math>. Carcinoid tumours of the colon and rectum do not contain the hydroxylase or decarboxylase enzymes and so do not form 5HTP or 5HT. A small number of patients have ‘atypical’ tumours. These patients excrete large amounts of 5HTP and 5HT in their urine. It is believed that these tumours, usually bronchial, are deficient in dopa decarboxylase and cannot convert 5HTP to 5HT so that the former is secreted into the bloodstream. Some of the 5HTP is converted to 5HT and subsequently to 5HIAA in extrarenal sites; some is decarboxylated in the kidneys and excreted into the urine as 5HT, and some escapes decarboxylation and is excreted directly into the urine. Thus, patients with atypical carcinoid tumours have$$$

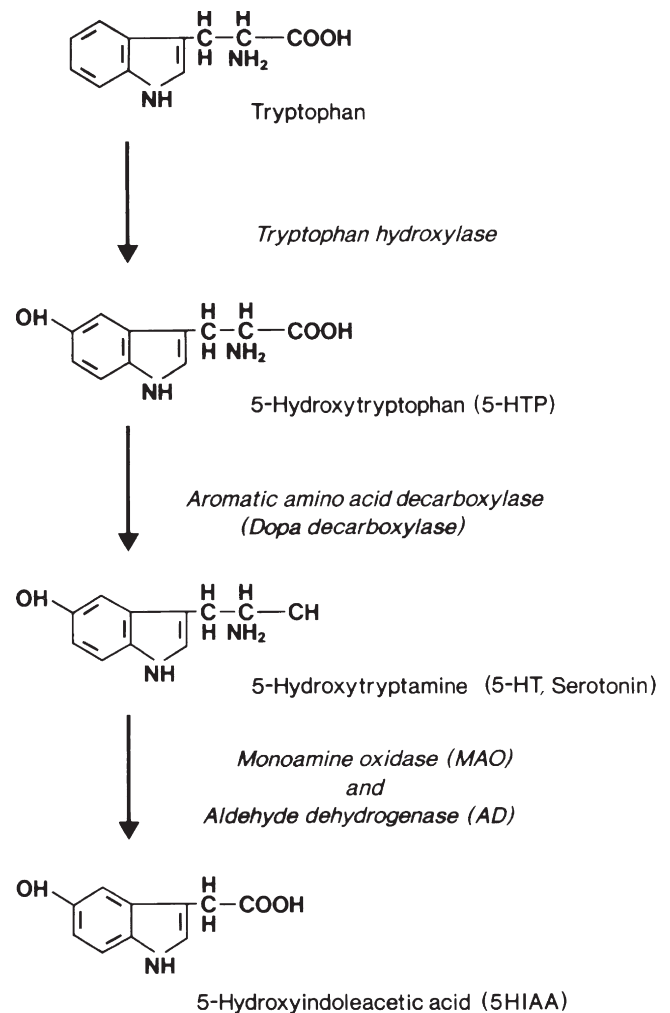


FIGURE 41.1 ■ Pathway of synthesis and metabolism of 5-hydroxytryptamine.

a marked increase in 5HTP and 5HT excretion and only a moderate increase in 5HIAA excretion.

Even in those patients whose tumours produce predominantly 5HTP, urinary 5HIAA constitutes 50–60% of total urinary 5-hydroxyindoles and increased values are found in almost all patients.

Laboratory investigation

The most sensitive indicator of the turnover of serotonin and its metabolites is measurement of 5HIAA in a 24h urine sample. A number of dietary substances can interfere in the measurement of 5HIAA and should be avoided for three days prior to and during the collection of the urine sample (see Box 41.2). The test is not completely specific and a negative result is to be expected in patients with non-secretory tumours and those of the hindgut that do not produce 5HT.

Chromogranin A (CgA) is a 48 kDa protein distributed in dense core granules of neuroendocrine cells, the highest plasma concentrations being found in patients with carcinoid tumours. It has been reported to have a sensitivity of 75–85% and a specificity of 84–95% in the diagnosis of carcinoid syndrome. It is not dependent upon

BOX 41.2 Substances that can interfere with the measurement of 5-hydroxyindole acetic acid

Serotonin-rich foods

- Bananas, aubergine, plums, pineapples, kiwi fruit, walnuts, hickory nuts, pecans, avocados

Artefactual increase in 5-HIAA

- Guafanesin (glycerol guaiacolate) contained in some cough medicines, paracetamol, fluorouracil, methysergide, naproxen, caffeine

Artefactual decrease in 5-HIAA

- Salicylates, L-DOPA, methyl dopa, aspirin, phenothiazines, adrenocorticotrophic hormone

serotonin secretion and is therefore particularly useful in patients with non-secretory or atypical carcinoids and as its concentration reflects tumour burden, it can be used as a marker for assessing response to treatment. False positive results can be seen in hepatic and renal failure, atrophic gastritis, proton pump inhibitor therapy, and inflammatory bowel disease. Prostatic cancers, which may contain a significant neuroendocrine component, myeloma, exercise, trauma and hypertension may also increase CgA concentrations.

Diagnostic imaging

In isolated carcinoid tumours, where surgical excision is a management option, anatomical localization may be useful. Conventional imaging modalities include computed tomography (CT), magnetic resonance imaging (MRI) and positron emission spectroscopy (PET). A more specific option is the use of somatostatin receptor scintigraphy, e.g. using ^{111}In -DTPA-pentetreotide (OctreoscanTM), which binds to somatostatin receptors on the tumour.

Treatment

Circumscribed tumours may be removable by surgical resection. For patients with significant hepatic metastatic disease, radiofrequency ablation or hepatic embolization, alone or in combination with intra-arterial chemotherapy (chemoembolization), while not curative, may provide significant palliation. Somatostatin analogues have been used effectively in the management of the symptoms of carcinoid syndrome.

MULTIPLE ENDOCRINE NEOPLASIA

Multiple endocrine neoplasia (MEN) syndromes are disorders characterized by pathological hyperfunction of two or more endocrine organs. They have been classified into two distinct disorders: MEN type 1 (MEN1) and MEN type 2 (MEN2) with MEN2 further subdivided into three main variants (see Box 41.3).

Two distinct genetic defects contribute to tumorigenesis in MEN syndromes: in MEN1, inactivation (loss of function) of a tumour suppressor gene is thought

BOX 41.3 Multiple endocrine neoplasia syndromes

Multiple endocrine neoplasia type 1

- Parathyroid adenoma
- Gastroenteropancreatic neuroendocrine tumours
- Anterior pituitary tumours

Multiple endocrine neoplasia type 2

MEN 2A

- Medullary thyroid carcinoma
- Pheochromocytoma
- Parathyroid disease

MEN 2B

- Medullary thyroid carcinoma
- Pheochromocytoma
- Absence of parathyroid disease
- Mucosal neuromas and intestinal ganglioneuromatosis
- Marfanoid habitus

MEN2A variants

- MEN2A with cutaneous lichen amyloidosis or Hirschsprung disease

Medullary thyroid carcinoma only

to be responsible, while MEN2 is caused by overactivation (gain of function or overexpression) of a proto-oncogene. A tumour suppressor gene restrains cell proliferation and tumours are stimulated by inactivating mutations or deletions in such a gene, which render the gene product either absent or nonfunctional. In contrast, a proto-oncogene (the normal unmutated version of an oncogene) can be converted to an oncogene, which can cause cell proliferations or deregulated cell growth when overexpressed.

Multiple endocrine neoplasia type 1

Multiple endocrine neoplasia type 1 (MEN1) typically becomes manifest after the first decade of life, with most men and women developing symptoms in the fourth and third decades, respectively. Typically, MEN1 tumours appear two decades earlier than isolated endocrine tumours.

Parathyroid disease

Primary hyperparathyroidism (HPT) is the most frequent endocrinopathy in MEN1 and is the most common reason for the disease to come to the attention of physicians. It occurs in 90% of affected individuals between 20 and 25 years of age, rising to nearly 100% by the age of 50 years. However, MEN1 itself is rare and accounts for only 2–4% of patients with HPT. There is no sex difference in MEN1 prevalence compared with a ratio of 3:1 females:males in sporadic hyperparathyroidism. Hyperparathyroidism is often found during the second decade of life during screening of immediate family members of an index patient with proven MEN1. The

pathological features are those of diffuse or asymmetrical hyperplasia, with all four glands being involved (although perhaps not all to the same degree) or with multiple adenomas. The investigation and management of HPT is discussed in Chapter 6.

Gastroenteropancreatic neuroendocrine tumours

Gastroenteropancreatic (GEP) neuroendocrine tumours are the second most common tumours in MEN1, with some 60% of patients being affected. They usually present between the second and fifth decades, unless diagnosed earlier by screening. Approximately half of the tumours are gastrinomas. The Zollinger–Ellison syndrome (ZES) is severe, intractable, multiple and recurrent peptic ulcer disease caused by a gastrin secreting tumour, which can be located in the duodenum or pancreas. The majority of gastrinomas associated with MEN1 are in the duodenum, where they are often small (<0.5 cm in diameter) and multiple. Since gastrin secreting cells are not normally found in the duodenum or pancreas, gastrinomas in these sites should be regarded as ectopic and potentially malignant, irrespective of their histological grade. The clinical syndrome does not differ from that seen with non-MEN1 gastrinomas. The extreme hypersecretion of gastric acid may be associated with inactivation of pancreatic lipase, resulting in fat malabsorption and steatorrhoea. Hyperparathyroidism in MEN1 can exacerbate hypergastrinaemia. The diagnosis can be made by demonstrating increased gastric acid secretion with simultaneously elevated plasma gastrin concentrations. Secretin infusion (2 U/kg) may cause augmented gastrin production in those patients whose basal values are equivocal. Excellent treatments are now available for the medical management of peptic ulcer disease using histamine H₂ receptor antagonists or proton pump inhibitors.

Other GEP NETs include insulinomas, vasoactive intestinal peptide secreting tumours (VIPomas) and glucagonomas. Their clinical features are discussed further in Chapter 12.

Chromogranin A is useful, particularly as a marker of midgut tumours, in the investigation of GEP NETs. Cocaine- and amphetamine-regulated transcript (CART), is a 116 amino acid peptide widely distributed in nervous and endocrine tissues and its physiological roles may include regulation of feeding and response to psychological stress. It is produced by various islet cell tumours and is of particular use as a marker in the investigation of NETs of pancreatic origin. Measurement of a screening panel of gut hormones may also be useful.

Pituitary tumours

The true prevalence of anterior pituitary tumours in MEN1 is unclear. Between 30% and 50% of patients with MEN1 develop anterior pituitary tumours. Whilst some are non-functional, approximately 60% secrete prolactin, 25% secrete growth hormone and 5% secrete ACTH. Occasionally, excess secretion of GH and cortisol

in MEN1 is the result of ectopic secretion of GHRH and ACTH, respectively, e.g. by a pancreatic islet or carcinoid tumour. It is important to identify such patients in order to ensure appropriate therapy.

The investigation and management of pituitary tumours in patients with MEN1 is similar to that for other pituitary tumours and is discussed in detail in Chapter 18.

Foregut carcinoid tumours

Carcinoid tumours occur more frequently with MEN1 than in the general population; some 10% of MEN1 patients are affected. In contrast to sporadic carcinoid tumours, which are predominantly derived from the midgut and hindgut, MEN1 carcinoid tumours are usually derived primarily from the foregut. Foregut carcinoids rarely secrete serotonin, peptide hormones or calcitonin and are usually considered as clinically non-functional.

Adrenal tumours

Adrenal cortical lesions are common in MEN1: 20–40% of patients are affected, mostly with bilateral tumours. However, the majority are non-functional, clinically silent and rarely require treatment. Cushing and Conn syndromes are rare in MEN1.

Tumourigenesis in MEN1

MEN1 is the result of an inactivating mutation in the *MEN-1* gene, a tumour suppressor gene located on chromosome 11q13. It consists of 10 exons and encodes a 610-amino acid nuclear protein termed menin. Menin interacts with diverse groups of transcription factors and coregulators, including JunD, suggesting a role in gene transcription, DNA replication and cell cycle control. Genetic mapping studies indicate somatic loss of heterozygosity in accord with the ‘two hit’ hypothesis. The first hit is a genetic mutation rendering the subject heterozygous for the *MEN-1* mutant gene and predisposed to tumour development. A somatic inactivation of the unaffected allele then occurs (second hit) leading to the development of MEN1-associated tumours.

Diagnosis of MEN1

Occasionally, MEN1 may occur without a recognized family history. Some may be patients whose parents died before developing any manifestations of MEN1: a few may have MEN1 caused by two somatic mutations as described above. Sometimes it is difficult to make the distinction between sporadic tumours and MEN1, but an earlier onset and tumour multiplicity in the same organ increase the likelihood of MEN1. The dominant mode of inheritance, the occurrence of asymptomatic gene carriers and the risk of developing malignant disease make it essential to establish carrier status in all relatives whenever the diagnosis of MEN1 has been made in one family member.

Genetic screening. The identification of the *MEN-1* gene (locus 11q13) has opened the possibility of genetic testing. However, over 1300 distinct germline mutations have been identified and there is a lack of genotype–phenotype correlation. Thus, genetic testing can be time consuming, arduous and expensive and a definitive result does not negate the requirement for further biochemical screening of carriers. However, for those found not to be carrying the mutant gene, there is no further need for surveillance or family screening.

Biochemical screening. Biochemical screening continues to play an important role in carrier ascertainment whenever genetic testing is not possible or has failed to detect germline mutations (see below) and is important for tumour surveillance in those known to have the affected gene.

The MEN1 syndrome rarely develops before the age of five or after the age of 70, and so screening should be performed every three years after the age of five until the age of 70 and at longer intervals thereafter. Over 95% of affected individuals will have been identified by the fourth decade. The measurement of serum calcium is a simple, cheap and reliable screening test; additional biochemical investigations include measurement of serum prolactin, gastrin (fasting) and CgA. Screening of patients with apparently sporadic pancreatic endocrine tumours for evidence of MEN1 is probably justified, especially in those with gastrinomas or insulinomas.

Surveillance of MEN1 patients and carriers

Once mutant *MEN1* gene carriers have been identified by biochemical or genetic testing, they should be monitored rigorously to detect the development of new tumours or tumour recurrence. The age-related penetrance is virtually zero below the age of five years, rising to >50% by 20 years, and >95% by 40 years. Surveillance should include a regular history and examination focused on the clinical features of MEN1-associated tumours in combination with biochemical tests and imaging (Table 41.1). Although periodic screening for endocrine tumour expression in MEN1 appears to reduce morbidity and mortality as a result of early tumour recognition, this has not been proven.

Multiple endocrine neoplasia type 2

Multiple endocrine neoplasia type 2 (MEN2) involves two or more of medullary thyroid carcinoma (MTC), pheochromocytoma and parathyroid tumours (see Box 41.3). Medullary thyroid cancer is a universal feature of MEN2 and is often the first tumour to present. MEN2A is the most common subtype, accounting for 80–90%. Patients with MEN2B, in addition to endocrine abnormalities, have a characteristic marfanoid habitus often with scoliosis and pes cavus. The presence of mucosal neuromas is characteristic (see Fig. 41.2). Some kindred appear to have a third form of MEN2, isolated familial MTC, but care should be taken that other features of MEN2 are not overlooked.

Germline activating mutations of the *RET* (rearranged during transfection) proto-oncogene, which is located on chromosome 10 (10q.11.2), have been identified in all three types of MEN2.

Diagnosis. RET germline mutation testing now forms the basis for carrier identification in MEN2 families. Medullary thyroid carcinomas secrete calcitonin, measurement of which serves as a useful tumour marker to assist in assessment of the response to therapy and for the detection of persistent or recurrent disease. Plasma carcinoembryonic antigen (CEA) concentration may be raised in medullary thyroid cancer and prove a useful additional tumour marker.

Before the establishment of DNA analysis for family screening for MEN2, measurement of calcitonin was often performed for this purpose, either on a



FIGURE 41.2 ■ Typical features of a patient with MEN2B. (Photograph courtesy of Professor A G McGregor)

TABLE 41.1 Surveillance for multiple endocrine neoplasia type 1

Procedure	Condition screened for	Frequency	Age initiated (years)
Glucose (fasting) and insulin	Pancreatic NETs	Yearly	5
Insulin-like growth factor 1	Pituitary tumour	Yearly	5
Prolactin	Pituitary tumour	Yearly	5
Calcium and PTH	Parathyroid tumour	Yearly	8
Chromogranin A, CART	Pancreatic NETs	Yearly	<10
Gastrin	Pancreatic NETs	Yearly	20
Brain MRI	Pituitary tumour	Every 3–5 years (or guided by biochemistry results)	5
Abdominal CT/MRI	Pancreatic NETs	Every 3–5 years (or guided by biochemistry results)	20

NETs, neuroendocrine tumours; PTH, parathyroid hormone; CART, cocaine and amphetamine-related transcript; CT, computed tomography; MRI, magnetic resonance imaging.

basal sample or in response to stimulation with pentagastrin or calcium. Every effort should be made to identify gene carriers among the relatives of patients known to have MEN2 in order to offer prophylactic thyroidectomy against MTC. Strong genotype–phenotype correlation exists in MEN2, which can guide the timing of surgery and the screening protocol for optimal detection of parathyroid tumours and pheochromocytomas.

Imaging. Ultrasonography of the thyroid may be useful. However, when invasion to surrounding structures or distant metastases are suspected, other imaging modalities such as MRI or CT scanning are necessary. ^{99m}Tc-technetium-sestamibi scan is a more sensitive technique for the localization of metastases within the neck and chest. Scintigraphic imaging using ¹²³I- or ¹³¹I-anti-CEA monoclonal antibodies is a promising development for the detection of occult disease and in selecting candidates for repeat neck exploration.

Treatment. If a mutation is found, surgical removal of the thyroid is recommended to prevent tumour development. Postoperatively, patients should be treated with suppressive doses of L-thyroxine. Since calcitonin gene expression is activated in inflamed or septic tissue, calcitonin concentration may remain elevated for 3–6 months after thyroid surgery. It is therefore prudent to delay calcitonin measurement for three months after thyroidectomy. Early and prophylactic total thyroidectomy has probably lowered the mortality from MTC to <5%, well below the cancer-associated mortality in MEN1. RET tyrosine kinase inhibitor drugs are currently being developed and may have a role in the treatment of MEN2-related malignancies in the future. Follow-up of MTC is by measurement of calcitonin and CEA.

Surveillance. Genotype–phenotype correlation in MEN2 may be useful in guiding strategies for surveillance. However, a suggested scheme is that after the age of eight years PTH and calcium should be measured for detection of parathyroid tumours and for detection of pheochromocytoma plasma free metanephrines and/or urinary fractionated metanephrines should be measured, with imaging using MRI or CT every 3–5 years or if biochemistry is abnormal. Patients with familial MTC require only periodic monitoring with imaging if biochemistry is abnormal.

Other familial syndromes associated with multiple endocrine neoplasia

Patients presenting with an endocrine tumour should have a full family history taken, looking for evidence of an inherited syndrome. Single, unilateral and late onset tumours are more likely to be sporadic. In contrast, a familial cause is more likely in a patient who presents early, particularly if there is a family history or if there are multiple tumours or associated cutaneous features.

A further type of MEN (MEN4) has been described caused by mutations in cyclin-dependent kinase inhibitor genes. The exact phenotype is, as yet, poorly defined but includes parathyroid, and sometimes pituitary, tumours and other endocrine features.

Pheochromocytoma is predominately a sporadic disorder. However, in 25% of patients, pheochromocytoma may be familial or associated with a number of autosomal dominant hereditary syndromes. Examples of such syndromes are neurofibromatosis type 1 (NF1), von Hippel–Lindau syndrome and paraganglioma syndrome 1, 3 and 4. The cardinal features of NF1 are neurofibromas and dermal café-au-lait skin pigmentation. Other features include skeletal manifestations, vascular stenoses and a variety of endocrine tumours including pheochromocytoma, somatostatin-producing carcinoid tumours and medullary thyroid cancer. The causative gene encodes a GTP-activating protein, neurofibromin. Von Hippel–Lindau syndrome is a multisystem cancer syndrome associated with retinal and cerebral haemangioblastomas, renal cysts and renal cell carcinomas, pheochromocytomas and islet cell tumours.

Paragangliomas (PGL) are tumours derived from the sympathetic and parasympathetic nervous system. About half are familial. The sympathetic-associated PGL arise from the adrenal medulla or from the sympathetic ganglia. They are generally functional and secrete excess catecholamines. Hereditary PGL are characterized by the development of highly vascularized, slowly growing non-chromaffin tumours arising in the parasympathetic ganglia. Parasympathetic-derived PGL usually develop in the head and neck, most commonly at the bifurcation of the carotid artery, and are usually non-functional. Intraabdominal and thoracic catecholamine-secreting PGL are currently being referred to as pheochromocytomas.

Paragangliomas syndromes have been genetically characterized as PGL 1, 3 and 4 and are caused by mutations in the succinate dehydrogenase (*SDH*) subunit D, C and B genes, respectively (*SDHD*, *SDHC* and *SDHB*); the fourth subunit coded by the *SDHA* gene is not associated with hereditary PGL.

METABOLIC CHANGES IN MALIGNANCY

Introduction

Metabolic changes may herald the clinical onset of malignant disease and are often the cause of considerable morbidity and mortality in patients with cancer. During the rapid growth phase of a neoplasm, the nutritional and metabolic state of the host may become modified in favour of the neoplasm. The tumour-related metabolic changes may be a direct consequence of the presence of the tumour, or be secondary to secretory products of the tumour. These may either be normal secretory products of the cells of origin or substances not normally produced by them. Such metabolic alterations may, on occasions, be useful in the diagnosis and management of malignancy, that is, they act as

tumour markers (see Chapter 42). In this section, some of the metabolic consequences of malignant disease are reviewed.

Paraneoplastic syndromes

A paraneoplastic syndrome is one that is attributable to a neoplasm, not as a direct result of local or metastatic presence of the tumour cells, but owing to substances secreted from them or autoantibodies produced against them. Such disorders may broadly be divided into neurological, humoral and other paraneoplastic syndromes.

Neurological paraneoplastic syndromes

Various neurological paraneoplastic syndromes exist. Some have a classic presentation, e.g. Lambert–Eaton, syndrome, which resembles myasthenia gravis and is associated with small cell carcinoma of the bronchus. Some paraneoplastic syndromes appear to have an autoimmune basis; thus Lambert–Eaton syndrome is associated with the presence of anti-voltage gated calcium channel (VGCC antibodies). The majority of neurological paraneoplastic syndromes are thought to have an autoimmune basis, with an immune response being directed against antigens expressed by the tumour. The resulting antibodies then cross react with components of the nervous system. These antibodies can be detected in blood. In a patient with a neurological presentation, the presence of certain well characterized onconeural antibodies (anti-Ho, Yo, Ma2, CRMP-5, amphiphysin and Ri) enables the diagnosis of a paraneoplastic syndrome to be made whether or not a neoplasm has been detected. Neurological paraneoplastic syndromes and associated tumours and antibodies are discussed further in Chapter 36.

Humoral paraneoplastic syndromes

Humoral paraneoplastic syndromes arise from the secretion of substances from a tissue that does not normally produce them. The term ‘ectopic hormone production’ is used to refer to hormones secreted from sites that are not their physiological origin, in amounts sufficient to cause clinical effects. The ectopic secretion of such hormones tends not to be subject to the usual mechanisms of endocrine regulation and characteristic responses to dynamic function testing may be useful in diagnosis. The hormone itself may sometimes circulate in a molecular form distinct from that secreted from the eutopic source. Treatment of the tumour leads to resolution of the endocrine effects, but these reappear if the tumour recurs, accompanied by a demonstrable concentration of the hormone in serum. Some tumours are particularly associated with such ectopic hormone production, e.g. small cell lung cancer. Various theories have been developed in an attempt to explain the phenomenon of ectopic hormone production by tumours. One theory suggests that some fundamental change occurs at the genomic level that allows novel gene expression for that tissue. Thus, neoplastic change would cause certain genes to be switched on and others to be

switched off. Another theory suggests that the neoplastic cell is derived from a stem cell that was capable of expressing the gene at an early developmental stage but is later suppressed, and that following neoplastic transformation, the cell undergoes dedifferentiation and regains some of its developmental properties. It has also been proposed that some widely expressed genes that are transcribed but not translated under normal conditions may, due to neoplastic transformation, be amplified and translated due to the action of different promoters.

Adrenocorticotropin

Ectopic secretion of adrenocorticotropin (ACTH) can occur with a variety of tumours, but is particularly associated with small cell carcinoma of the bronchus. Its clinical features and diagnosis are discussed in Chapter 18. The most common cause of ectopic ACTH production is the expression of the proopiomelanocortin (POMC) gene by tumour tissue. The *POMC* gene encodes for the 31-kDa precursor of ACTH, β -endorphin and melanocyte stimulating hormone. The principal source of POMC-derived peptides in the body is the pituitary gland; however, POMC immunoreactivity and POMC mRNA can be found in almost all tissues. The mRNA species produced in non-pituitary tissue is shorter than that expressed in pituitary tissue, which is thought to be due to the action of different promoters. One theory is that many tumours express the *POMC* gene, but what determines whether the tumour becomes an ectopic ACTH-secreting tumour is the switching of activity to the more conventional pituitary promoter. Excessive ACTH secretion can also be the result of ectopic secretion of corticotrophin-releasing hormone.

Vasopressin

The synthesis of vasopressin, resulting in the paraneoplastic syndrome of inappropriate anti-diuresis is particularly associated with small bronchial carcinoma, although can occur with other tumours (see Table 41.2). It may present acutely with severe symptomatic, hyponatraemia. This condition is discussed further in Chapter 4. Vaptans, agents which antagonize the action of vasopressin via blockade of the vasopressin receptor, have been used in management.

PTH-related peptide

Hypercalcaemia complicating malignancy is common. Mechanisms include prostaglandin-mediated bone resorption, osteoclast activating factors, e.g. tumour necrosis factors, and the action of PTH-related peptide (PTHrP), a protein with N-terminal amino acid homology with PTH. PTH-related peptide is not detected by most modern PTH assays which measure ‘intact’ PTH, hence the finding of a low PTH with a low phosphate in a hypercalcaemic patient is suggestive of hypercalcaemia secondary to PTHrP. Hypercalcaemia as a feature of malignancy is discussed further in Chapter 6.

TABLE 41.2 Hormones arising from tumours derived from tissue not classically associated with their secretion

Hormones	Associated tumour types
Adrenocorticotrophic hormone, lipotrophin, melanocyte-stimulating hormone, endorphin, enkephalin, other pro-opiomelanocortin fragments	Small cell carcinoma of bronchus Adenocarcinoma of bronchus Bronchial carcinoid tumour Thymic carcinoid tumour Medullary thyroid carcinoma Islet cell carcinoma of pancreas Pheochromocytoma Gut, prostate, neurogenic and parotid tumours
Atrial natriuretic peptide Arginine vasopressin	Small cell carcinoma of bronchus Small cell carcinoma of bronchus Pancreatic carcinoma Breast carcinoma Thymic tumour
Calcitonin (clinically silent)	Small cell carcinoma of bronchus Bronchial carcinoid tumour
Calcitonin gene-related peptide Corticotrophin releasing hormone Erythropoietin	Small cell carcinoma of bronchus Bronchial carcinoid tumour Small cell carcinoma of bronchus Uterine fibroma Cerebellar haemangioblastoma Pheochromocytoma Ovarian tumour
Gastrin releasing peptide	Small cell carcinoma of bronchus Medullary thyroid carcinoma
Glucagon	Non- β islet cell tumour Anaplastic lung carcinoma Renal adenocarcinoma
Growth hormone	Bronchial carcinoma Gastric carcinoma
Growth hormone releasing hormone	Pancreatic islet cell tumour Pheochromocytoma
Human chorionic gonadotrophin	Renal and bladder carcinomas Breast carcinoma Prostatic carcinoma Melanoma Gynaecological carcinomas Head and neck carcinomas Lymphoma
Human placental lactogen Insulin-like growth factor 2 Neurotensin	Lung and testicular tumours Mesodermal and mesenchymal tumours Pancreatic endocrine tumours Carcinoid tumour
Parathyroid hormone-related peptide	Squamous cell carcinoma of head and neck, oesophagus cervix, bronchus Breast and ovarian carcinomas Carcinoma of pancreas Renal and bladder carcinomas Multiple myeloma Histiocytic lymphoma
Prolactin	Anaplastic lung carcinoma Small cell carcinoma of bronchus Renal adenocarcinoma
Renin	Renal adenocarcinoma Adenocarcinoma of pancreas Small cell carcinoma of bronchus Adenocarcinoma of lung Ovarian carcinoma
Somatostatin	Adrenocortical carcinoma Small cell carcinoma of bronchus Duodenal somatostatinoma Extra-adrenal paraganglionoma Pheochromocytoma Medullary thyroid carcinoma
Vasoactive intestinal polypeptide	Bronchial carcinoid tumour Ganglioneuroblastoma Pheochromocytoma Non- β islet cell tumour Small cell carcinoma of bronchus

Tumour-induced osteomalacia

Tumour-induced osteomalacia (TIO) is a paraneoplastic syndrome resulting from the secretion of fibroblast growth factor 23 (FGF-23) by mesenchymal tumours. There may be a relatively long history of musculoskeletal symptoms such as bone pain and muscle weakness and, in children, rickets and growth failure may be observed. Classic biochemical features are normal plasma calcium and PTH concentrations and hypophosphataemia with renal phosphate wasting. Plasma 1,25-dihydroxyvitamin D concentration is low and FGF-23 elevated. The only effective treatment for TIO is resection of the tumour. Medical management is with phosphate supplements and calcitriol.

Other paraneoplastic syndromes and features of malignant disease

Haematological sequelae. Anaemia is the most common haematological abnormality encountered in malignant disease. It may have various causes with differing laboratory findings (Table 41.3). Clinical and laboratory features are discussed further in Chapter 27; examples include anaemia of chronic disease, and folate deficiency resulting from altered metabolism or requirements due to the presence of the tumour. Autoimmune haemolytic anaemia can occur, most commonly in patients with B cell lymphoproliferative disorder, and is thought to be caused by an immune response evoked to antigens common to both the tumour cells and red blood cells. In contrast to idiopathic autoimmune haemolytic anaemia, response to corticosteroid therapy is uncommon. Pure red cell aplasia has been described mainly with thymomas, but also in association with adenocarcinomas, squamous cell carcinomas, anaplastic tumours and, rarely, with lymphoproliferative diseases. The mechanism is thought to be a T cell-mediated phenomenon. Microangiopathic haemolytic anaemia is a serious complication of certain malignancies and occurs when red cells are damaged by passage through blood vessels that have been distorted by either tumour or fibrin deposits. The blood film is characterized by the presence of schistocytes and polychromasia and there is biochemical evidence of haemolysis (increased plasma LDH activity,

reduced haptoglobin concentration and haemosiderinuria). In contrast, some tumours, particularly of the kidneys, may be associated with erythrocytosis: this is thought to be the result of prostaglandin-mediated erythropoietin activity.

Hyperuricaemia. In certain patients with malignancy, predominantly those with leukaemia and lymphoma in whom there is a rapid turnover of cells, or massive cell lysis caused by cytotoxic agents (tumour lysis syndrome), hyperuricaemia can occur. Radiotherapy can produce a similar clinical scenario. Other features of the tumour lysis syndrome include hyperkalaemia, hyperphosphataemia and hypocalcaemia.

The increased filtration of urate and the increasing acidity and concentration of the tubular fluid leads to precipitation and renal obstruction. Adequate hydration is important prior to combination chemotherapy in those at risk and it is standard practice to administer allopurinol, a xanthine oxidase inhibitor, to reduce urate formation. Rasburicase (recombinant urate oxidase) catalyses the conversion of urate to allantoin, which is five to ten times more soluble than urate, and is helpful in reducing plasma urate in severely affected patients. Rasburicase should be avoided in patients with glucose 6-phosphate dehydrogenase deficiency as one of the major by-products of the enzymatic reaction is hydrogen peroxide, which may precipitate haemolysis. Preservation of samples at 4°C, or stabilization with perchloric acid, has been recommended prior to analysis of urate in patients receiving rasburicase, as degradation of uric acid can continue in collection tubes, resulting in a spuriously low urate concentration.

A number of other characteristic paraneoplastic syndromes exist. These include immune-mediated renal disease, thrombosis associated with antiphospholipid antibodies and finger clubbing and hypertrophic pulmonary osteoarthropathy attributed to the action of platelet-derived growth factor.

Cancer cachexia

Weight loss is common in cancer patients and in many, is the presenting feature. Cancer cachexia is a hypercatabolic state characterized by loss of body weight associated

TABLE 41.3 Haematological indices of anaemia in malignant disease

Cause	Hb	Mean cell volume	Mean cell Hb	Mean cell Hb conc.	Reticulocytes	Film	Marrow
Iron deficiency	↓	↓	↓	↓	Normal/↑	Hypochromic microcytic	Reduced iron stores
Folate and/or vitamin B ₁₂ deficiency	↓	↑	↑	Normal	Normal/↑	Macrocytic	Megaloblastic
Anaemia of chronic disease	↓	Normal	Normal	Normal	↓	Normocytic	Sideroblastic
Marrow infiltration	↓	Normal/↑	Normal/↑	Normal/↑	Normal/↓	Normocytic ± malignant cells	Leukoerythroblastic

Hb, haemoglobin.

with a reduction in adipose tissue and muscle mass and loss of appetite. Weight loss can cause severe morbidity and decrease in quality of life and may be associated with an adverse prognosis and increased susceptibility to side-effects during treatment of the malignancy itself. The aetiology of the weight loss is complex: it appears to involve a combination of increased catabolism and decreased anabolism. The physical presence of the tumour may also contribute to weight loss, e.g. by causing obstruction in the gastrointestinal tract.

The regulation of appetite involves both central and peripheral mechanisms, which are integrated in the hypothalamus. There is no evidence that hypothalamic dysfunction per se is responsible for the cachexia of malignant disease. However, it has been suggested that malignant tissue causes anorexia by the secretion of biologically active molecules that depress feeding by interfering with central control mechanisms, e.g. serotonin from carcinoid tumours and bombesin from small cell bronchial cancers. Cytokines, particularly TNF α and IL-6, are thought to be important in the loss of appetite. They arise from tumour cells and as part of the host immune response; they are able to cross the blood-brain barrier and interact with mechanisms affecting appetite. Iatrogenic factors related to the treatment of cancer can also contribute to weight loss. Chemotherapy often causes nausea and vomiting, and radiotherapy can have short-term effects on the gut

(e.g. mucositis) and cause long-term complications, e.g. enteritis that lead to malabsorption.

Changes in metabolism

Although reduced energy intake is common to both starvation and cancer cachexia, there are marked differences between the metabolic changes seen in the two states (see Table 41.4).

Many patients with cancer are mildly hypermetabolic with a resting energy expenditure greater than normal. However, only a fraction of the increased metabolic rate can be accounted for by the tumour tissue itself. Suggested mechanisms for cancer cachexia are shown in Figure 41.3.

TABLE 41.4 Metabolic changes in cancer cachexia and starvation

	Starvation	Cachexia
Acute phase response	No	Yes
Appetite	Increased	Decreased
Metabolic rate	Decreased	Increased
Skeletal muscle mass	Maintained	Decreased
Adipose tissue	Decreased	Decreased
Liver size	Decreased	Increased

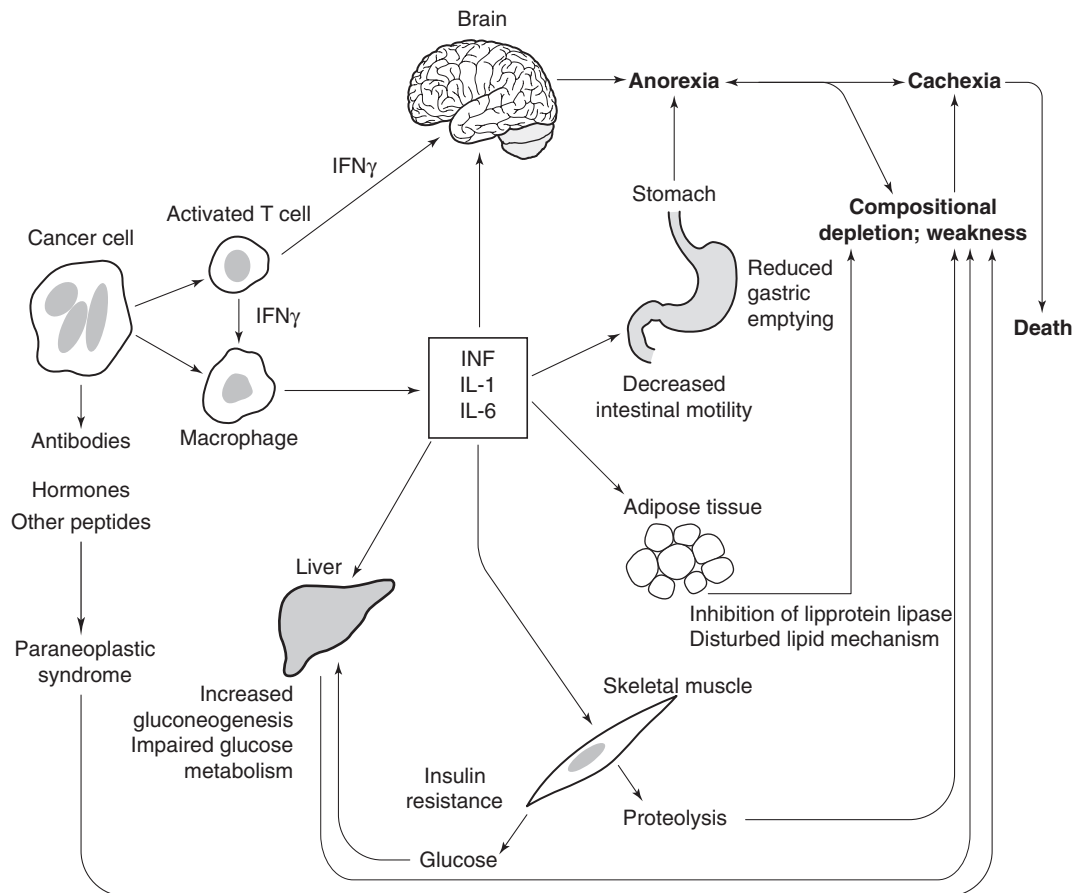


FIGURE 41.3 ■ Possible mechanisms of cancer cachexia.

Many tumours produce energy via anaerobic glycolysis, which results in a net loss of ATP, there is partial uncoupling of oxidative phosphorylation and futile cycles may be activated leading to further energy wastage, e.g. nonesterified fatty acids released from adipose tissue are immediately re-esterified. Free fatty acids are oxidized, even when other energy sources are available, and there is increased protein turnover.

Marked derangements of carbohydrate metabolism are a feature of malignancy. They include abnormal glucose tolerance and hyperglycaemia, hypoglycaemia, lactic acidosis, increased glucose turnover and increased glucose transport into tumour cells. Tumour cells have an increased rate of glucose transport, which may be related to an increase in the number of membrane bound glucose transporters. The main energy source for many tumour cells is anaerobic metabolism of glucose, which results in the production of lactic acid. The lactic acid is then reconverted to glucose in the liver via the Cori cycle. This process also consumes energy, since glycolysis produces only two moles of ATP per mole of glucose, whereas gluconeogenesis consumes six moles of ATP. However, decreased hepatic capacity to utilize lactic acid may lead to its accumulation (lactic acidosis).

Glucose intolerance is a common feature of malignancy, affecting some 60% of cancer patients. Occasionally, frank hyperglycaemia is present. There is insulin resistance (at least with respect to effects on carbohydrate metabolism), although insulin receptor binding appears normal. Fasting insulin concentrations are also normal or slightly decreased but there is a decreased response to both endogenous and exogenous insulin. The picture is further complicated by other factors that may be present in cancer patients, i.e. starvation and malnutrition, sepsis, bed rest, dietary effects and drugs.

Abnormal lipid metabolism is a feature of some malignancies. Some tumours utilize lipids in preference to glucose as a major energy source, with mobilization of free fatty acids from the body's fat stores and a consequent rise in their plasma concentration. Malignant cells may show abnormal lipid composition, particularly with respect to the cholesterol and phospholipid contents of their cell membranes and may synthesize unusual lipids. For example, hepatocellular carcinomas may produce 2-methylolate. Loss of fat mass is facilitated by a tumour-produced lipid mobilizing factor. This appears to be a proteoglycan similar in structure to zinc α_2 -glycoprotein, which acts to sensitize adipose tissue to lipolytic stimuli by increasing production of cyclic AMP within adipocytes.

Patients with malignancy often have a wasting syndrome, where both adipose tissue and lean tissue are lost. The muscle wasting is due to increased protein catabolism as well as decreased synthesis. Proteolysis in cachexia is brought about via increased activity of a number of proteolytic pathways. The major mechanism for proteolysis in cancer cachexia is the ATP ubiquitin-dependent proteolytic pathway initiated in response to cytokines such as TNF α . Lysosomal cathepsins and the calcium/calpain pathway are also involved in proteolysis.

Treatment

Although giving nutrition support to patients with cancer cachexia would seem logical, there is little evidence that provision of energy to patients, via either enteral or parenteral routes, results in weight gain or significant improvement in outcome. Pharmacological management of cachexia has included use of appetite stimulants and drugs intended to counteract catabolism or stimulate anabolism. Corticosteroids have been used to stimulate appetite. Their effect is short lived and is not usually accompanied by weight gain. The progesterone analogue megestrol acetate has been shown to be more effective as an appetite stimulant. Various anabolic agents have been tried in cachexia. Growth hormone has proven efficacy in animal models but its use in clinical settings is associated with increased mortality, possibly owing to diversion of amino acids and energy to skeletal muscle away from the acute phase response. Examples of the use of drugs intended to treat cancer cachexia by modification of the immune response, include thalidomide and pentoxifylline, which both inhibit production of TNF α , and eicosapentaenoic acid, found in fish oil, which can reduce lipolysis produced by lipid mobilizing factor. None of these has been shown to be consistently effective.

ENDOCRINE SEQUELAE OF TUMOURS AND THEIR TREATMENT

With the development of successful therapies for specific malignancies, populations of treated patients are surviving for near-normal life spans. Long-term sequelae are becoming more evident as these survivors are being followed-up (Table 41.5). This section focuses on effects on somatic growth and reproductive capacity.

Effects on somatic growth

Somatic growth may be profoundly suppressed in a child affected with, for example, acute lymphoblastic leukaemia. This effect may be mediated by therapy aimed at destroying the malignant cells, including cranial irradiation with subsequent hypopituitarism, direct radiotherapy to bony structures that determine final height, for example the spine, or the administration of pharmacological doses of glucocorticoids and intensive chemotherapy.

The hypothalamus is more susceptible to radiation damage than the pituitary. Growth hormone secretion appears to be particularly vulnerable to the effects of radiation, and, while doses of 18–50 Gy may lead to decreased growth hormone secretion alone, higher doses (>60 Gy) can cause panhypopituitarism through damage to both the hypothalamus and pituitary. Effects of chemotherapy and glucocorticoids can have significant effects on skeletal development, affecting both final height and bone mineral density. The effects of chemotherapeutic agents on chondrocytes, the extracellular matrix and osteocytes are wide ranging, but evidence for direct effects on chondrocytes' proliferation and

TABLE 41.5 Endocrine sequelae of treatment of malignant disease

Organ affected	Endocrine sequelae
Hypothalamo-pituitary	
Combination chemotherapy: cyclophosphamide, vinca alkaloids, melphalan, cisplatin	Syndrome of inappropriate antidiuresis (SIADH)
Radiation therapy	Panhypopituitarism
Glucocorticoids (pharmacological doses)	Suppression of hypothalamo-pituitary-adrenal axis
	Suppression of growth hormone secretion
	Sick euthyroid syndrome
GnRH analogues	Hypogonadotrophic hypogonadism
Somatostatin	Suppression of growth hormone secretion
Antiemetics (dopaminergic antagonists)	Hyperprolactinaemia
Thyroid	
Systemic illness	Sick euthyroid syndrome
Radiation therapy	Hypothyroidism
Bleomycin, vinblastine, cisplatin (BVP)	Retarded thyroid hormone clearance
Glucocorticoids (pharmacological doses)	Sick euthyroid syndrome
Pancreas	
Combination chemotherapy: vincristine, L-asparaginase, streptozotocin, cyclophosphamide, mitomycin C, 5-fluorouracil	Reduced insulin secretion
Adrenals	
5-Fluorouracil, enzyme inhibitors	Decreased steroidogenesis and hypoadrenalism
Gonads	
Chemotherapy	
Definite: L-phenylalanine mustard, busulphan, cyclophosphamide	Reduced spermatogenesis and suppressed Leydig cell function in the male
Probable: methotrexate, 5-fluorouracil, 6-mercaptopurine	Suppression of ovarian cyclicity and ovarian steroidogenesis in the female
Radiation therapy	Hypergonadotrophic hypogonadism (primary ovarian or testicular failure)
GnRH analogues	Hypogonadotrophic hypogonadism

ossification is limited; however, disruption of growth plate activity will result in skeletal growth disturbance. Chemotherapy seems to enhance the growth-suppressing effects of radiotherapy when they are given in combination, possibly by enhancing the effects of radiation on the hypothalamo-pituitary axis. It has been hypothesized that chemotherapy effects may relate to suppression of IGF1 synthesis in the liver and/or its effects on growth plates.

For growth hormone to effect growth optimally, normal rhythms of secretion must be maintained. This will only occur if there is normal hypothalamo-pituitary function. The secretion of growth hormone may be affected directly as a result of radiation-induced damage to the hypothalamus or pituitary, or secondarily, for example, by depression, which can accompany any severe illness and is associated with impaired growth hormone secretion. Thyroid hormone and glucocorticoids are necessary for the efficient biological activity of growth hormone; the secretion of either may be decreased following radiotherapy to the hypothalamo-pituitary unit and lead indirectly to impairment of growth. Pharmacological doses of glucocorticoids suppress growth hormone secretion. Additionally, they cause a degree of functional hypogonadism (hypogonadotrophic hypogonadism). If gonadal steroidogenesis is suppressed, the pubertal growth spurt will not occur and final height will be compromised. Glucocorticoids and gonadal steroids ultimately bring about long bone epiphyseal fusion. Their pharmacological use may hasten this fusion and so prevent the attainment of the

genetic potential for final height. Low-dose cranial irradiation can also result in early puberty with long bone fusion before attainment of final height.

Strategies to promote growth in these patients can be divided into two groups: either the use of growth hormone replacement in growth hormone deficient individuals, or the use of GnRH analogues to delay puberty in those in whom it is occurring early. Earlier and more aggressive use of growth hormone in those who are growth hormone deficient has resulted in improvements in final height, with safety data suggesting that this does not lead to an increased risk of tumour recurrence. Caution should be exercised, however, as longer term follow-up data are not yet available to determine whether there is an increase in secondary neoplasms. The use of GnRH analogues in combination with growth hormone therapy is more controversial: only some of the few studies suggest an improvement in final height.

Reproductive consequences of therapy

Disease itself, or radiotherapy or chemotherapy directed against populations of actively dividing cells, may influence reproductive potential. This effect may be mediated by a direct action on the germ cell population or be an indirect effect mediated by suppression of hypothalamic GnRH or pituitary gonadotrophin secretion. Specific guidelines are available that may help to predict doses of radiotherapy that may be tolerated, or drug combinations that are less likely to cause depletion of the germ cell pool. Thus, a chemotherapeutic regimen of adriamycin,

bleomycin, vinblastine and dacarbazine (ABVD) is less likely to cause infertility than the alternative regimen of nitrogen mustard, oncovin, procarbazine and prednisone (MOPP) in the treatment of Hodgkin disease, a fact that may influence the choice in a young patient who is anxious to subsequently have a family. As the germ cell pool is greater in younger than older women, the risk of ovarian failure as a result of treatment increases with age. Use of GnRH analogues to suppress oocyte maturation via the hypothalamo–pituitary axis and to make the germinal epithelium quiescent (a premenarche state) has shown some success in protecting ovarian function against chemotherapy in human and animal studies. Other strategies that may be used include oophoropexy (an operation to relocate the ovaries anatomically so that they are shielded from the radiotherapy field) or sperm banking prior to therapy. Freezing ovarian tissue pre-chemotherapy with transplant post-chemotherapy has been shown to be successful at restoring function, including fertility, although this process carries with it the potential risk of the re-introduction of tumour cells.

Radiotherapy can affect the uterus, with a decrease in its volume and changes in its musculature and vascular supply. Embryos or oocytes can be frozen prior to therapy, but success rates for subsequent viable pregnancy are limited by the freeze–thaw process. Furthermore, the time factor involved in stimulating a population of oocytes for retrieval prior to commencement of therapy often militates against this, and ovarian stimulation itself has the potential to increase tumour growth in oestrogen-sensitive tumours.

Some tumours, such as prostatic or breast cancers, are actively stimulated by a particular endocrine milieu. Thus, androgens stimulate the growth of prostatic cancer, while oestrogens may encourage the growth of breast cancer. Direct modulation of the endocrine axis may, in itself, constitute therapy for these particular malignancies. Thus growth of prostatic cancer may be suppressed by orchidectomy or by suppression of the hypothalamo–pituitary–testicular axis by a depot injection of a GnRH analogue. This seeming paradox may be explained by the fact that chronic stimulation with GnRH is associated with downregulation of pituitary receptors and inhibition of gonadotrophin release. Alternatively, the

target organ effect of androgens may be blocked by the administration of an anti-androgen such as cyproterone acetate or flutamide. In women, oestrogen effects may be opposed by the administration of an anti-oestrogen, such as tamoxifen, or the use of an aromatase inhibitor such as 4-hydroxyandrostenedione, which blocks the penultimate step in oestrogen biosynthesis.

These patients have specific concerns not just about their reproductive capacity, but also about the likelihood of transmission of the malignancy to their offspring or the risk of damage to their germ cells, making fetal malformation more likely than in the normal population. With the exception of the heritable syndromes, there is no evidence of an increased likelihood of malignancy or an increased risk of fetal malformation in progeny of treated patients at the present time.

CONCLUSION

Tumours can have profound metabolic effects, even when very small. These can be due to the physical presence of the tumour, to the products of its secretion (whether expected on the basis of its tissue of origin or not) or to the body's response to the presence of the tumour. Metabolic derangements can also occur as a side-effect of treatment, whether this be surgical or with radiotherapy or chemotherapy.

Further reading

- Chong WH, Molinolo AA, Chen CC. Tumour-induced osteomalacia. *Endocr Relat Cancer* 2011;18:1253–77.
- Gozzard P, Maddison P. Which antibody and which cancer in which paraneoplastic syndrome? *Postgrad Med J* 2011;87:60–70.
- Melmed S, Polonsky KS, Larsen PR et al. editors. *Williams' textbook of endocrinology*. 12 ed. Philadelphia: Saunders Elsevier; 2011.
- Ramage JK, Ahmed A, Ardill J et al. Guidelines for the management of gastroenteropancreatic (including carcinoid) tumours (NETS). *Gut* 2012;61:6–32.
- Tisdale MJ. Mechanism of cancer cachexia. *Physiol Rev* 2009;89:381–410.
- Wassif WS, Moniz CF, Friedman E et al. Familial isolated hyperparathyroidism: a distinct genetic entity with an increased risk of parathyroid cancer. *J Clin Endocrinol Metab* 1993;77:1485–89.
- Wassif WS, Farnebo E, Teh BT et al. Genetic studies of a family with hereditary hyperparathyroidism–jaw tumour syndrome. *Clin Endocrinol* 1999;50:191–96.

Tumour markers

Catharine M. Sturgeon

CHAPTER OUTLINE

INTRODUCTION 821

Evaluation of the clinical utility of tumour markers 821

Tumour marker requests and the responsibilities of the clinical laboratory 825

TUMOUR MARKERS IN THE MANAGEMENT OF SPECIFIC CANCERS 827

Bladder cancer 827

Breast cancer 827

Cervical cancer 829

Choriocarcinoma 829

Colorectal cancer 829

Gastric cancer 831

Gastrointestinal stromal tumours (GIST) 831

Germ cell tumours 832

Gestational trophoblastic neoplasia 834

Hepatocellular carcinoma (primary liver cancer) 835

Lung cancer 836

Melanoma 837

Neonatal and paediatric tumours 837

Ovarian cancer 838

Pancreatic cancer 840

Prostate cancer 840

Testicular cancer 841

Thyroid cancer 842

Cancers of unknown primary 842

SUMMARY 842

INTRODUCTION

Tumour markers are substances (often proteins, enzymes or hormones) that are present in body fluids or tissues and whose measurement provides information about the presence, progression or remission of tumours. They may be tumour-derived (produced within the tumour by malignant or stromal cells) or tumour-associated (produced by non-malignant cells as a metabolic consequence of tumour presence). Some are tumour-specific (produced by cancerous but not normal tissue), while others are present in normal tissue but are produced at higher concentrations in body fluids or malignant tissue from cancer patients. A few are organ-specific but many are produced by a variety of different cancers.

Histopathological identification of tumour markers expressed at the tumour cell surface and detectable in biopsy specimens can provide both diagnostic and prognostic information, with genetic tests becoming increasingly important in predicting which patients are most likely to respond to costly new therapies. Secreted tumour markers, present in blood or other body fluids, are measured quantitatively, often by immunoassay. While knowledge of circulating marker concentrations may assist in diagnosis and prognosis, their most important clinical use is in monitoring the success of treatment.

This chapter focuses primarily on the properties and clinical applications of the tumour marker tests that usually fall within the remit of clinical biochemistry laboratories (Tables 42.1 and 42.2), with the aim of enabling

readers to provide advice about which tumour marker tests are most likely to be helpful, to recognize which results require immediate action, and to develop a high-quality interpretative tumour marker service. Before considering these aspects, however, it is useful to review some recent developments relevant to the introduction of tumour markers into routine clinical practice. Abbreviations used in the chapter for the names of individual tumour markers are explained in Tables 42.1 and 42.2.

Evaluation of the clinical utility of tumour markers

Although hundreds of potential tumour markers have been investigated, with more than 200 000 papers describing them, the number that contribute significantly to the management of cancer patients is remarkably small, as is evident in Tables 42.1 and 42.2. Historically, over-enthusiastic reporting of small, poorly designed studies of new markers has led to disillusionment when the results have not been confirmed in larger studies or in other centres. Together with the difficulties inherent in comparing different studies reported in non-standard form, this has made objective assessment of the utility of many tumour markers very difficult and has meant that the recommendations made by expert panels about their clinical use tend to be conservative.

The three critical factors that should be considered when evaluating a tumour marker are its clinical utility,

TABLE 42.1 Characteristics of commonly requested tumour markers

Tumour marker	Biochemical properties	Molecular weight	Main clinical applications	Limitations of use
Alkaline phosphatase	Phosphohydrolase	Variable	Raised activities associated with the presence of liver and or bone metastases	Raised in pregnancy and in children, and in non-malignant liver and bone disease
α -Fetoprotein (AFP)	Glycoprotein, 4% carbohydrate; considerable homology with albumin	~70 kDa	Diagnosis and monitoring of hepatocellular carcinoma, hepatoblastoma and germ cell tumours Prognosis of germ cell tumours	Raised in pregnancy and neonates, and in benign liver disease, gastric and other gastrointestinal (GI) tract tumours, e.g. oesophageal and pancreatic
Cancer antigen 125 (CA125)	Mucin identified by monoclonal antibodies OC125 and M11 Developed from serous cystadenocarcinoma cell line OVCA 433	~200 kD	Monitoring ovarian carcinoma Determination of the risk of malignancy index (RMI) for ovarian carcinoma	Raised in patients with pleural effusions, ascites or free fluid in the pelvis, and congestive cardiac failure Also raised in benign renal and liver disease, and in other adenocarcinomas Mildly raised during menstruation and in the first two trimesters of pregnancy Variably raised in endometriosis. May be increased by peritoneal trauma, e.g. postoperatively
Carcino-embryonic antigen (CEA)	Family of glycoproteins, 45–60% carbohydrate	~180 kD	Monitoring colorectal adenocarcinomas	Absent or low concentrations in poorly-differentiated tumours, poor sensitivity for malignancy in early disease, slightly raised in benign renal, liver, lung and GI disease Also raised in other malignancies, including breast, gastric, lung, mesothelioma, oesophageal and pancreatic Smokers may have higher concentrations than non-smokers Contamination with saliva may markedly increase concentrations
Human chorionic gonadotrophin (hCG)	Glycoprotein hormone consisting of two non-covalently bound subunits (α and β); α -subunit similar to LH, FSH, TSH; β -subunit considerable homology with LH	~36 kD	Diagnosis, prognosis and monitoring of germ cell tumours and gestational trophoblastic neoplasia	Raised in pregnancy and in other tumours, e.g. lung Use of cannabis can transiently elevate serum concentrations in patients with germ cell tumours
Lactate dehydrogenase (LDH)	Enzyme of the glycolytic pathway	Variable	Diagnosis, prognosis and monitoring of germ cell tumours To monitor a wide range of malignancies, including haematological	Elevated in cardiac and benign liver disease and some anaemias of non-malignant cause
Paraproteins	Monoclonal immunoglobulins	Variable	See Chapter 30	
Prolactin	Pituitary hormone.	~22 kD but high molecular forms also exist	See Chapter 18	
Prostate-specific antigen (PSA)	Glycoprotein; member of the kallikrein family with serine protease activity Circulates as free enzyme or complexed to α_1 -antichymotrypsin (measurable) or α_2 -macroglobulin (not detected by most immunoassays)	~30 kD (free enzyme)	Diagnosis, risk assessment and monitoring of prostatic carcinoma Lower concentrations of free PSA relative to complexed PSA (i.e. free: total ratio) found in prostatic cancer as compared with benign prostatic hypertrophy	Essentially organ-specific but elevated in both benign and malignant prostatic disease, serum concentrations increase with age Transiently elevated in urinary tract infections, prostatitis, acute retention and following catheterization or other manipulation of the prostate (e.g. prostatic biopsy, transurethral resection of the prostate) Specimens for free PSA should be analysed within 4 h of sampling or stored frozen until assayed

TABLE 42.2 Characteristics of less readily available tumour markers

Tumour marker	Biochemical properties	Molecular weight	Main clinical applications	Limitations of use
Calcitonin Cancer antigen 15-3 (CA15-3), BR 27.29	32 amino acid peptide Mucin (MUC-1 glycoprotein peptide) identified by monoclonal antibodies	~3.5 kD >250 kD	See Chapter 19 Monitoring breast cancer	Raised in benign liver, breast and ovarian disease and in other malignancies, e.g. lung, colon, ovary
Cancer antigen 19-9 (CA19-9)	Glycolipid carrying the Lewis blood group determinant	~1000 kD	Monitoring pancreatic carcinoma following curative resection	Not detected in subjects negative for the Lewis blood group determinant Raised in obstructive jaundice, cholestasis, cirrhosis, hepatitis, pancreatitis and non-malignant GI disease Contamination with saliva may markedly increase concentrations
Catecholamines Chromogranin A	Biogenic amines Member of the granin family of acid secretory glycoproteins	~0.2 D ~49 kD	See Chapter 38 Monitoring neuroendocrine tumours	Raised in renal, liver and cardiac failure
CYFRA 21-1	Fragments of cytokeratin 19	~30 kD	Monitoring lung carcinoma	Contamination with saliva may markedly increase concentrations
Gut hormones, e.g. vasoactive intestinal peptide (VIP), pancreatic polypeptide (PP), somatostatin, gastrin	Small peptide hormones	Variable	See Chapter 12	
Human epidermal growth factor receptor-2 (HER-2 or c-erb2)	Transmembrane glycoprotein encoded from <i>HER-2/neu</i> oncogene	185 kDa	Predicting response to trastuzumab (Herceptin®)	
Human epididymis protein 4 (HE4)	Product of the <i>WFDC2 (HE4)</i> gene that is over-expressed in patients with ovarian carcinoma	~25 kD	Monitoring ovarian cancer. Potentially an aid in diagnosis as part of the risk of malignancy algorithm (ROMA) Under evaluation	Higher concentrations in postmenopausal women
Inhibin A (α - β_A) Inhibin B (α - β_B)	Heterodimeric glycoproteins composed of an α - and a β -subunit. There are several forms of the β -subunit	32 kDa	Monitoring of ovarian granulosa cell tumours and testicular Sertoli and Leydig tumours	
β_2 -Microglobulin	Component polypeptide chain of the HLA antigen complex	~11 kD	See Chapter 30	
Neuron specific enolase (NSE)	Dimer of the enzyme enolase	~87 kD	Monitoring small cell lung carcinoma, neuroblastoma and neuroendocrine tumours	Haemolysis must be avoided Specimens should be separated within 1 h
Oestrogen receptor (ER)	Nuclear transcription factor	~5 kD	Predicting response to endocrine therapy in breast cancer Part of the histological assessment of tissue specimens	
Prostate cancer gene 3 protein (PCA3)	Protein product of <i>PCA3</i> gene in urine of prostate cancer patients	~86 kD	Potentially an aid to diagnosis of prostate cancer particularly in biopsy-negative men Under evaluation	
Placental alkaline phosphatase (PLAP)	Heat-stable isoenzyme of alkaline phosphatase	~86 kD	Monitoring of germ cell tumours (seminomas)	Raised in smokers

(Continued)

TABLE 42.2 Characteristics of less readily available tumour markers (Continued)

Tumour marker	Biochemical properties	Molecular weight	Main clinical applications	Limitations of use
Progesterone receptor	Nuclear transcription factor	A form: ~4 kD B form: ~120 kD	Predicting response to endocrine therapy in breast cancer Part of the histological assessment of tissue specimens	
Squamous cell carcinoma antigen (SCC)	Glycoprotein sub-fraction of tumour antigen T4	48 kD	Monitoring squamous cell carcinomas	Contamination with saliva or skin may markedly increase concentrations
S-100	Polypeptide homodimer		Melanoma	
Thyroglobulin (Tg)	Glycoprotein dimer of two identical subunits	670 kDa	Monitoring differentiated thyroid cancer	Raised in benign thyroid disease Potential assay interference from autoantibodies
Tissue polypeptide antigen (TPA)	Fragments of cytokeratin 8, 18 and 19	~22 kD	Monitoring bladder and lung carcinoma	Contamination with saliva may markedly increase concentrations

TABLE 42.3 Examples of applications of tumour marker measurements at different stages of diagnosis and treatment

Application	Examples
Assessment of risk	PSA: used with or without digital rectal examination to assess risk of prostate cancer and determine the need for biopsy
Screening	HCG: screening of women who have had a previous molar pregnancy and who are at high risk of developing choriocarcinoma
Differential diagnosis	CA125: together with menopausal status and ultrasound findings, contributes to calculation of the risk of malignancy index in the differential diagnosis of women with pelvic masses
Prognosis: prediction of relapse or progression: in primary disease	CEA: measurement at three-monthly intervals following curative surgery in patients to assess need for further treatment
in metastatic disease	Thyroglobulin: following ablation of the thyroid, increasing serum concentrations suggest an alternative site of production
Prognosis: prediction of response to therapy: in primary disease	Oestrogen and progesterone receptors: their presence or absence in breast cancer tissue determines whether endocrine therapy is likely to be effective
in metastatic disease	AFP, hCG and LDH: in patients with germ cell tumours, concentrations of these markers are used to assess prognosis
Monitoring course of disease	
To detect relapse in patient with no evidence of disease post-therapy	AFP, hCG and LDH used in the follow-up of patients treated for germ cell tumours
To follow detectable disease	CA125 used to monitor ovarian cancer patients

the magnitude of the benefit of its use and its reliability. Tumour marker measurement can provide information at multiple stages of diagnosis and treatment (Table 42.3) but, until appropriate studies have been performed, it can be difficult to know how best to use a tumour marker in particular circumstances. Having identified a specific use (i.e. utility), the clinical value (i.e. magnitude) of using the tumour marker for that application needs to be assessed, by evaluating the difference in outcome between marker-positive patients and marker-negative patients, ideally in a randomized controlled trial designed for the purpose. Clinical precision and accuracy (i.e. reliability) then need to be established, since the marker will only be useful if results are reproducible. In this context, it is important to note that, although it is necessary to demonstrate a statistically significant difference in tumour

marker concentrations between patient groups to show that a marker may have potential utility, this is not in itself sufficient evidence that a marker is of clinical benefit (i.e. that it should be used), but merely suggests that the differences observed are not likely to occur by chance. It is also essential to ensure that biochemical analysis of the marker is reliable and reproducible, that assays are standardized and that their analytical and clinical performance is tested objectively in appropriately designed and well-conducted studies.

The Tumour Marker Utility Grading System (TMUGS), developed some years ago, provides a useful framework for such evaluation and it is encouraging that the recommendations made therein have been implemented for some analytes, e.g. by improvements in the standardization of immunohistochemical assays for HER-2 and in

the analytical accuracy and equimolarity of serum assays for PSA. The TMUGS also describes levels of evidence for grading the clinical utility of tumour markers, an approach that has been adopted by the US National Academy of Clinical Biochemistry (NACB), which has developed Laboratory Medicine Practice Guidelines for the use of tumour markers in the clinic. Other initiatives include development of Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK), which provide guidance about study design, pre-planned hypotheses, patient and specimen characteristics, assay methods and statistical analysis methods and complement the broader statements on Consolidated Standards of Reporting Trials (CONSORT) and Standards for Reporting of Diagnostic Accuracy (STARD). These should help to encourage improved design and publication of tumour marker studies in the future.

Tumour marker requests and the responsibilities of the clinical laboratory

Reasons for requesting tumour markers

Knowledge of tumour marker results can contribute useful clinical information relating to different aspects of patient management (Table 42.3). For use in screening and diagnosis, the 'perfect' tumour marker would be absent in all healthy subjects (100% specificity), would be raised in all patients with a single tumour type (100% sensitivity) and its serum concentration would accurately reflect tumour size. Unfortunately, the 'perfect' marker does not exist and the extent to which currently available markers meet this ideal is very variable. Consequently, the predictive value of a positive or negative result also varies and is highly dependent on the population considered (see Chapter 2 for detailed discussion).

There is increasing interest in improving the specificity of some markers (e.g. AFP, CA125 and PSA) by using serial measurements to assess the rate of change in the concentration of the marker with time in individual patients. Such measurements may contribute to diagnosis, but in practice, are more likely to aid post-treatment monitoring, which is the most established clinical role for tumour markers. Whether such monitoring improves patient outcome depends on the availability of further treatment options, should increases in marker concentrations indicate progressive disease. If potentially curative treatment is available, as for example is frequently the case for patients with recurrent germ cell tumours, it is essential to identify patients with rising tumour marker concentrations promptly; failure to do so may constitute a critical clinical error. However, if alternative therapy is not available, knowledge of progressive disease before it is clinically evident may not benefit the patient and can have adverse psychological consequences, in which case it may be advisable to discontinue tumour marker monitoring. Serial monitoring is also desirable for patients with cancers that are unlikely to progress and for which treatment is not necessary or can be delayed: active surveillance (sometimes called 'watchful waiting') programmes, including serial PSA measurements, are appropriate for some men with indolent prostate cancers.

Choice of tumour marker test

In practice, the clinical biochemistry laboratory receives three types of tumour marker requests: those for patients with diagnosed malignancy who have already been referred to specialist centres; those for patients being investigated for suspected malignancy in secondary care, and those for patients presenting to their general practitioners. Requests in the first group are the most likely to be appropriate and are usually made to clarify the patient's diagnosis or to monitor response to treatment and/or detect recurrence.

Non-specialist users, whether in hospital or general practice, should be encouraged to consider carefully whether knowledge of a tumour marker result is likely to be helpful before requesting it. Requests such as 'tumour marker screen' or '?malignancy', particularly from emergency departments and other receiving units, should be actively discouraged. Since it is not usually practicable for requests to be scrutinized prior to analysis, laboratories should provide comprehensive advice about appropriate choice of tumour markers, as well as reminders of their limitations. Most tumour markers do not have sufficient sensitivity or specificity, particularly for early stage disease, to be regarded as diagnostic tests, although they may contribute to diagnosis. Importantly, whatever the malignancy or tumour marker, a result within the reference interval never excludes malignancy or progressive disease.

Requestors should be aware that the converse is also true. Raised tumour marker concentrations do not necessarily indicate malignancy, as they may be increased in a number of benign conditions. Increased concentrations may be associated with more than one tumour type, since, with few exceptions, tumour markers are not organ-specific. Clinical biochemists themselves should also think carefully before requesting any additional testing which may lead to a diagnosis of malignancy and before doing so should seek the agreement of the doctor managing the patient. General advice that can be readily disseminated to non-specialist users has recently been prepared by Pathology Harmony UK.

Pre-analytical requirements

Although the timing of blood sampling is not usually critical, a pre-treatment specimen is helpful when interpreting subsequent results. Specimens should always be taken before any investigative procedure, since some of these may cause transient releases of tumour markers into the circulation (e.g. increases of PSA following insertion of a urinary catheter or prostatic biopsy, of CA125 following abdominal surgery, and of CEA following colonoscopy). Possible conditions, which can transiently affect tumour marker concentrations (e.g. marked increases of PSA in men with active urinary tract infections or of CA19.9 in patients with cholestasis) should be excluded. Failure to recognize that misleadingly high results can be obtained if sampling times are inappropriate may cause undue distress to the patient, as well as decreasing confidence in laboratory testing. Some additional examples are listed in Tables 42.1 and 42.2.

TABLE 42.4 International standards (IS), reference preparations (IRP) and reference reagents (IRR) for tumour markers

Tumour marker	Code	Year established	Description
International standards and reference preparations			
AFP	IS 72/225	1972	Crude cord serum (50%): mass units
CA125	No International Standards yet available		
CA15-3			
CA19-9			
CEA	IRP 73/601	1973	CEA purified from liver metastases of primary colorectal cancer: arbitrary units
hCG	IS 07/634	2007	Highly purified urinary hCG: arbitrary units and molar units
hCG α	IS 75/569	1975	Arbitrary units
hCG β	IS 75/551	1975	Arbitrary units
International reference reagents			
hCG	IRR 99/688	2001	Highly purified: molar units
hCGn	IRR 99/642	2001	Nicked form of hCG: molar units
hCG α	IRR 99/720	2001	Highly purified urinary hCG α , dissociated from hCG: molar units
hCG β	IRR 99/650	2001	Highly purified, free from intact dimer: molar units
hCG β n	IRR 99/692	2001	Partially degraded hCG β : molar units
hCG β cf	IRR 99/708	2001	Molar units
PSA	IRR 96/670	2001	90:10 ratio of bound to free PSA: mass units
fPSA	IRR 96/668	2001	Purified free PSA: mass units
Thyroglobulin	CRM 457	1996	Purified thyroglobulin: mass units

Most currently available immunoassays are calibrated against the relevant standards where they exist. The lack of international standards for the CA antigens is a major hindrance to improved between-method comparability. hCG α , hCG α -subunit; hCG β , hCG β -subunit; hCGn, nicked hCG; hCG β n, nicked hCG β -subunit; hCG β cf, hCG β -core fragment. International Standards and Reference Reagents are available from the National Institute for Biological Standards and Control, Potters Bar, Herts, UK: http://www.nibsc.ac.uk/catalog/standards/preps/sub_endo.html

Analytical requirements

Manufacturers' instructions should always be followed when performing tumour marker measurements, with carefully defined internal quality control (IQC) and external quality assessment (EQA) procedures in place to monitor performance. Internal quality control and EQA specimens should closely resemble patient sera and be of clinically relevant concentrations, including those near important decision limits. Regular assessment of reproducibility and stability of results with time is particularly important for all tumour markers, as these are often monitored over long periods. Ensuring good reproducibility at low concentrations is critical where treatment may be instituted solely on the basis of a relatively small increase in tumour marker concentration, as is the case for AFP and hCG in germ cell tumours and for PSA following prostatectomy for prostate cancer.

Major international efforts are being directed towards improved comparability of tumour marker methods by encouraging accurate calibration against the relevant international standards where these exist (Table 42.4), by producing reference reagents to enable improved characterization of assays, by organizing collaborative workshops to identify the most clinically appropriate antibody specificities and by encouraging use of equimolar assays where relevant (e.g. PSA assays that recognize free and complexed PSA equally well). Nevertheless, the molecular heterogeneity of most tumour markers (as illustrated by the number of hCG-related molecules in Table 42.4) means that results obtained with different methods are not interchangeable. Clinical biochemists should be aware of the characteristics of the methods used in their own laboratory. They should also be familiar with the vulnerability of their methods to potential interferences such as the high-dose hook effect, specimen carry-over and interference from heterophilic or human anti-mouse antibodies (Box 42.1).

BOX 42.1 Potential causes of misleading results that are particularly relevant to tumour markers

High-dose hook effect

- Possible tumour marker concentrations range over several orders of magnitude. Protocols permitting identification of high-dose 'hooking' are essential to avoid reporting misleadingly low results, particularly in patients for whom markers are being measured for the first time. (Hook effects can be minimized by using solid-phase antibodies of higher binding capacity, by assaying specimens at two dilutions, or by using sequential assays that include a wash step.)

Specimen carry-over

- Potentially a problem whenever very high concentration specimens are assayed, so should occasionally be checked

Interference from heterophilic or human anti-mouse antibodies (HAMA)

- Falsely high or low results may be obtained for patient specimens containing anti-immunoglobulin G (IgG) antibodies capable of reacting with antibodies used in the assay. Presence of HAMA, frequently induced in cancer patients who have undergone treatment with mouse monoclonal antibodies for imaging or therapeutic purposes, may also give erroneous results. Such interference can be detected by re-assaying the specimen after treatment with a blocking agent (which is commercially available immobilized on tubes), by adding further non-immune mouse serum to the reaction mixture, or by re-assaying the specimen by a different method. An example of such interference is given in Table 42.8.

Reporting of tumour marker results

Cumulative and/or graphical reporting of serial results can identify trends in marker concentrations, which are generally more informative than single values, and may alert the laboratory to unexpected results (e.g. sudden changes) that require further investigation. Recording brief clinical information (e.g. 'postoperative'), preferably both in the laboratory computer and on any printed reports, can be helpful both in identifying results that are out-of-accord and when interpreting results.

In view of method-related differences in tumour marker results, it is recommended that the method used is stated on the clinical report. If there has been an intervening method change, it is highly desirable that the laboratory also indicates whether this is likely to have affected interpretation of the trend in results. Reference intervals specific to the method used should be provided, although for serial monitoring, the patient's own baseline provides the most important reference point for future results. Laboratory and clinical staff should engage in active dialogue about appropriate clinical decision points (e.g. when using PSA measurements to select patients for biopsy) and should know how these limits were derived. Laboratories can increase the value of their reports by adding short interpretative comments relating to the analytical results and preferably individually tailored to the requestor (e.g. omitting these for specialist users). Tumour marker half-lives, which are defined as the time to 50% reduction of circulating tumour marker concentrations following complete removal of tumour tissue, provide an important measure of the efficacy of therapy, particularly for germ cell tumours, and it may be helpful for the laboratory to calculate these for AFP and hCG (see p. 833 for further information). Providing advice about whether a change in marker concentration is likely to be significant or not is also helpful and should take into account both biological and analytical variation. A confirmed increase of 30%, or two serial increases of 20–25%, are often considered to be of clinical significance. Recommendations about the need for confirmatory specimens and the desirable frequency of tumour marker measurements are also helpful.

The clinical biochemist should identify urgent results that may be required for immediate patient management and ensure that these reach the relevant clinician promptly, telephoning results when appropriate. Results in this category include those which can be used to diagnose advanced disease in critically ill but treatable patients (e.g. AFP in hepatoblastoma; hCG in choriocarcinoma; AFP and hCG in non-seminomatous germ cell tumours; PSA in men with advanced prostate cancer that may respond to endocrine therapy). Provision of a proactive and high-quality tumour marker service helps to encourage good communication between laboratory and clinical staff, and is likely to facilitate both appropriate use of tumour marker tests and early identification of any results that are not in accord with the clinical picture. An example of a laboratory report meeting many of these requirements is shown in [Figure 42.1](#).

TUMOUR MARKERS IN THE MANAGEMENT OF SPECIFIC CANCERS

Regularly updated national guidelines on the management of the majority of tumours are now widely

available, often provided electronically on the world-wide web. Regional cancer networks are also a valuable source of information and have often developed modified versions of national guidelines tailored for local use. The optimal use of tumour markers for assessment of prognosis, monitoring treatment and detecting recurrent disease has been studied in most detail for choriocarcinoma and germ cell tumours, relatively rare diseases for which tumour marker measurements are mandatory for clinical management. Measurements of serum tumour markers also contribute significantly to the management of some of the more common malignancies (e.g. ovarian, colorectal and prostate), while they are less widely used in others (e.g. bladder, breast and lung). In the following sections, the extent to which tumour markers currently contribute to the clinical management of a number of important malignancies is briefly reviewed.

Bladder cancer

The most common symptom of bladder cancer is intermittent haematuria, which is present in 80–85% of patients. The majority of bladder cancers are transitional cell carcinomas but adenocarcinomas, squamous cell carcinomas and sarcomas also occur. In some patients, urine cytology is positive for tumour cells, but the diagnosis is usually established by cystoscopic evaluation.

Urine cytology is very effective in detecting high-grade bladder cancers but will miss the majority of papillary urothelial neoplasms of low malignant potential. Commercially available assays for two tumour markers present in urine have been approved by the United States Food and Drug Administration (FDA) for the detection of recurrent bladder cancer. The BTA (bladder tumour associated antigen) -Trak™ and -Stat™ tests detect complement factor H and related proteins, which are involved in the regulation of the alternative pathway of complement activation to prevent complement-mediated damage to healthy cells. A point of care version of this test is also available. The nuclear matrix protein 22 (NMP22™) test is a quantitative measure of the nuclear mitotic apparatus protein, a component of the nuclear matrix which is over-expressed in bladder cancer. Both these tests are more sensitive than cytology in detecting low grade bladder cancers but are less specific, and their high false positive rates limit their clinical application. Better specificity has been achieved with a fluorescence in situ hybridization assay (UroVysion™), which detects bladder cancer-associated aneuploidy of selected chromosomes and which has been approved by the FDA for screening patients for recurrent bladder cancer.

Breast cancer

Breast cancer is by far the most common cancer affecting women worldwide with approximately one million new cases diagnosed each year. The main presenting features in women with symptomatic breast cancer include a lump in the breast, nipple change or discharge and skin contour changes.

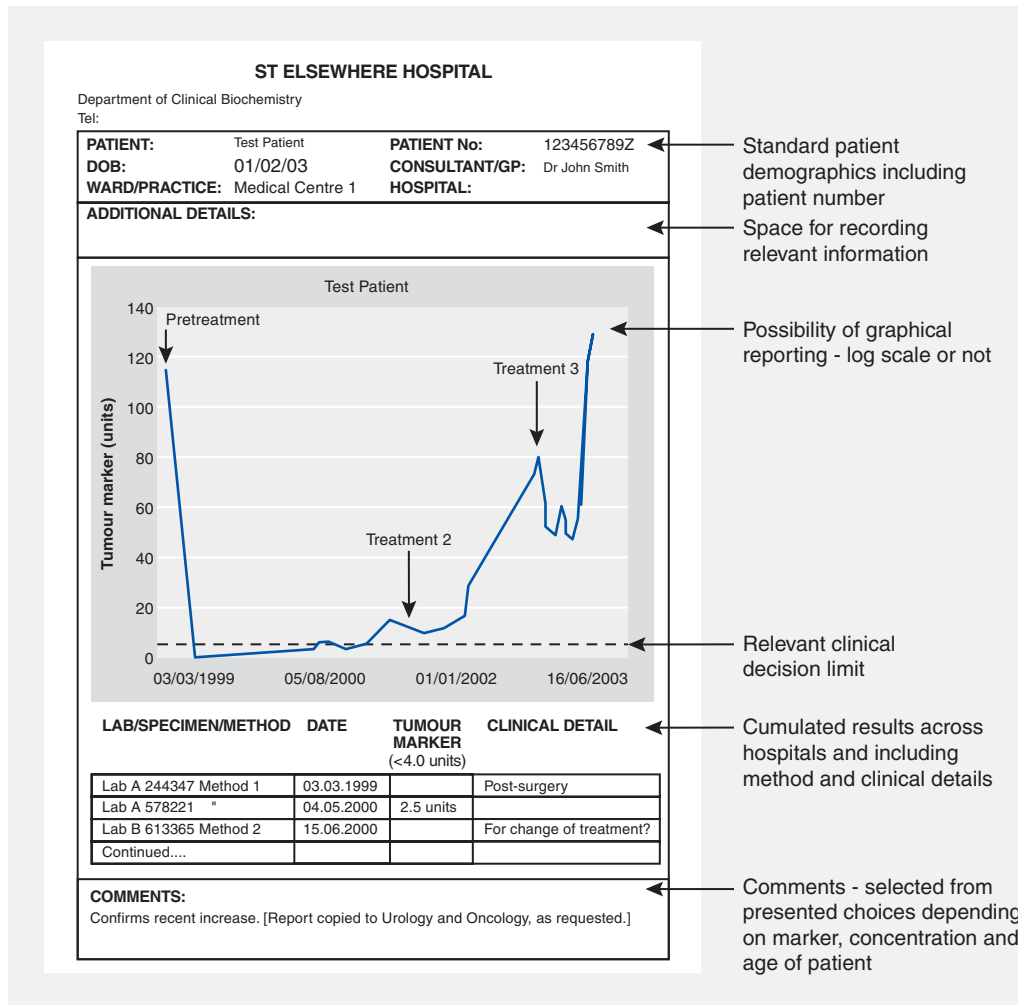


FIGURE 42.1 ■ Possible template for a clinical laboratory report for tumour markers that fulfil current reporting recommendations.

Screening and diagnosis

Currently available blood-based biomarkers are of no value in the early diagnosis of symptomatic or asymptomatic breast cancer, the latter being addressed by national screening programmes using mammography. Individuals who are at increased risk of breast cancer because they are carriers of the *BRCA1*, *BRCA2* or *TP53* genetic mutations and those with a strong family history of breast cancer (e.g. close relatives diagnosed with breast cancer at a young age) may be eligible for screening with magnetic resonance imaging (MRI), as is recommended in the UK by the National Institute for Health and Care Excellence (NICE). Definitive diagnosis requires biopsy and histopathology.

Prognosis

Measurement of oestrogen-binding receptors (ER) and progesterone-binding receptors (PR) in tumour biopsy tissue obtained at diagnosis is mandatory in order to determine the likely response to endocrine therapy of both early and advanced (metastatic) breast cancer. Early ligand-binding and enzyme-linked immunosorbent assay methods

for measuring ER and PR have been superseded by immunohistochemical assessment, which can be performed on paraffin-fixed tissue sections of the smaller tumours now detected by screening. Immunohistochemical or fluorescence in-situ hybridization tissue measurement of HER-2 (or its gene), a glycoprotein that controls cell growth and is amplified in ~30% of early-stage breast cancers, is now essential in all newly diagnosed patients with breast cancer. Patients with tumours that do not produce HER-2 are not likely to benefit from treatment with trastuzumab (Herceptin®), a humanized monoclonal antibody against HER-2, either administered on its own or in addition to other chemotherapy treatment. An assay is available for the measurement of the soluble shed form of HER-2 in serum and has potential value, both for prognosis and for monitoring trastuzumab therapy in patients with advanced breast cancer. However, the serum assay is not widely used.

Monitoring

CA15-3, a high molecular weight mucin, and similar MUC-1 based glycoprotein markers (e.g. BR27-29) may be used to monitor response to treatment, rising serum concentrations providing early indication of progression

in some patients. Their routine use is not currently recommended, since there is as yet no evidence that therapeutic intervention prior to radiological or clinical detection of recurrent tumour is beneficial. However, in individual patients, measurement of CA15-3 may help to determine whether bone symptoms are due to benign or malignant disease, as high concentrations are frequently associated with metastatic disease. CA15-3 may also be useful in monitoring response to therapy in advanced disease if there are no other indicators of response and if rising concentrations would be an indication to stop all but palliative treatment. As with other tumour markers, the low sensitivity of CA15-3 means that results within the reference interval do not exclude active disease or progression.

Cervical cancer

Worldwide, cervical cancer is the major cause of death from gynaecological cancer, with reported incidence rates in developing countries much higher than those in developed countries. Since cervical cancer progresses slowly from pre-invasive cervical intraepithelial neoplasia to invasive cancer, screening asymptomatic women with regular smears provides an effective means of early detection. The addition of human papilloma virus (HPV) testing appears to improve the effectiveness of screening as certain types of HPV are involved in the development of cervical cancer.

Women who have not been screened may present with symptoms of abnormal vaginal bleeding and, in advanced cases, pelvic pain and pressure symptoms relating to the bowel or bladder. Treatment of early stage disease usually requires radical hysterectomy and pelvic lymphadenectomy, with adjuvant radiation therapy if disease has spread to the lymph nodes. Surgery and concomitant chemoradiation or neoadjuvant chemotherapy followed by radical surgery are options for bulky disease. Measurement of plasma CEA and CA125 may have clinical utility in patients with cervical adenocarcinomas, which constitute 10–15% of all cervical cancers. The majority of cervical cancer patients (~85%), however, have squamous cell cervical carcinomas, for which the potentially most useful serum tumour marker is squamous cell carcinoma antigen (SCC), a serine protease inhibitor.

Screening and diagnosis

The low sensitivity and specificity of SCC, particularly for early stage disease, preclude its use in screening or diagnosis, which requires immunohistological evidence. Although 60% of patients with cervical cancer will have elevated concentrations of SCC at diagnosis, raised values are also found in other squamous cell carcinomas (e.g. lung, oesophagus and head and neck) as well as in benign diseases (e.g. psoriasis, eczema, sarcoidosis).

Prognosis

Some studies suggest that a raised concentration of SCC is an independent risk factor for poor survival but other studies contradict this. It has also been suggested

that pre-treatment SCC concentrations may be helpful in stratifying patients at greater risk of recurrence and therefore requiring more intensive therapy, but formal trials will be necessary to confirm this.

Monitoring

Various studies have confirmed that SCC shows a strong correlation with the clinical course of cervical cancer, with a lead time of up to 14 months in detecting progression prior to the onset of clinical symptoms, but whether earlier detection improves treatment outcome is not known. As only 10% of patients with recurrent disease can be cured and most patients with recurrent disease have clinical symptoms, until treatment options improve it is unlikely that SCC measurement will be widely adopted.

Choriocarcinoma

See Gestational trophoblastic neoplasia, below.

Colorectal cancer

Worldwide, colorectal cancer (CRC) is the third most common cancer, with more cases in the developed world than in Africa and Asia. Its incidence increases as populations adopt a western style diet. The risk of recurrence and subsequent death from CRC is closely related to the stage of the disease at the time of the primary operation. Although the treatment of colon and rectal cancer differ, they can be considered together in terms of tumour marker measurement.

Screening

Individuals at high risk of developing CRC (e.g. those with hereditary non-polyposis colon cancer (HNPCC, Lynch syndrome), familial adenomatous polyposis (FAP) or a strong family history of CRC (e.g. first degree relative with CRC diagnosed before age 45) should be referred to a clinical genetics unit for risk assessment and endoscopic screening. Mismatch repair gene mutations (e.g. *MLH1*, *MSH2*) predispose individuals to additional tumours, including those of the endometrium, ovary, genitourinary tract, small bowel and biliary tract. Individuals with FAP are at increased risk of developing duodenal tumours and thyroid tumours.

Early detection of CRC using faecal occult blood testing (FOBT) has been shown to reduce mortality in several randomized controlled trials. A number of countries now provide screening programmes for defined age groups (often >60 years old in the general population and at younger ages in high-risk groups) based on FOBT. Such screening is undertaken biannually in the UK, currently using a guaiac test, which detects the pseudoperoxidase activity of either intact or free haemoglobin. The test has many limitations, including relatively low clinical sensitivity and specificity for CRC; lack of specificity for human haemoglobin; vulnerability to interference from some foodstuffs and medications and difficulties associated with automating it. Screening using guaiac-based methods is therefore gradually being

replaced by faecal immunochemical tests, which detect the globin component of haemoglobin, can be used quantitatively with an adjustable cut-off concentration and are generally superior. All FOBTs lack specificity, and positive screens must be followed-up with colonoscopy. Markers based on DNA detection are potentially more specific than FOBT and may ultimately replace them, provided a clinically cost-effective panel of markers can be identified. At present, however, the cost and technical difficulties associated with these assays preclude their adoption.

Diagnosis

Although CEA is the most frequently used marker for CRC, depending on the cut-off point chosen, serum CEA will be raised in only 30–50% of CRC patients at the time of diagnosis. Its specificity is also low, since CEA concentrations can be raised in benign liver and kidney disease as well as in other malignancies (e.g. breast, gastric, lung, mesothelioma, oesophageal and pancreatic cancers), and may also be raised in smokers. Therefore, CEA cannot be used in isolation to diagnose even advanced CRC. However, raised concentrations can aid in diagnosis in certain clinical circumstances (e.g. indicating a high probability of malignancy in a frail elderly patient who cannot undergo invasive investigations) and confirmed markedly raised concentrations (serum CEA >40 µg/L) are suggestive of metastatic disease.

Prognosis and staging

The strongest predictor of recurrence of CRC is the histopathological stage of disease, as assessed using the Dukes staging system or one of its modified versions (Table 42.5), which take into account local tumour invasion, involvement of regional lymph nodes and presence of distant metastases. The five-year survival for patients with Dukes Stage A disease is >90% but that for those with Stage D disease is <10%.

TABLE 42.5 Pathological staging systems used for colorectal cancer

Extent of spread of malignancy	Pathological stage	
	Dukes	UICC/AJCC
Tumour found in the innermost lining of the colon only (carcinoma in situ)	–	0
Tumour beginning to spread but confined to the inner lining of the colon or rectum	A	I
Tumour extending more deeply into and/or through the colon or rectum, possibly invading nearby tissue	B	II
Invasion of tumour to regional lymph nodes	C	III
Metastasis to distant sites (most frequently lungs and liver)	D	IV

UICC, International Union Against Cancer; AJCC, American Joint Committee on Cancer.

High preoperative CEA concentrations are associated with poorer prognosis, while those within the reference interval are generally associated with better outcomes. An Expert Panel of the American Joint Committee on Cancer has suggested that CEA concentrations should be incorporated in the TNM staging system for CRC, but this proposal has not yet been formally adopted. Supporting this, a College of American Pathologists Expert Panel has ranked preoperative serum CEA as a Category I prognostic marker for CRC. In practice, pathological examination of resected tumour currently determines prognosis and the need for adjuvant therapy but preoperative CEA concentrations may provide additional prognostic information complementing histopathology in newly diagnosed CRC patients. Further information may also be obtained from new prognostic markers, e.g. microsatellite instability and gene expression profiling. Further evaluation of these is required prior to their routine implementation but it is likely that measurement of CEA will continue to be both simpler and cheaper.

Monitoring

Serum CEA is a reliable and validated marker for the surveillance and detection of disease recurrence in CRC. However, it is important to consider whether this would be likely to benefit the individual patient. Monitoring of the disease should be undertaken only if it were likely to lead to clinical action that would either improve the survival or quality of life of the patient, e.g. by early detection of recurrence that is amenable to treatment.

Meta-analyses of a number of trials confirm that intensive follow-up results in 20–30% reduction in all cause mortality, but variation in follow-up strategies and reporting makes it difficult to draw conclusions about the best combination of tests and their timing. Nevertheless, results suggest that inclusion of regular CEA measurements in intensive follow-up regimens is necessary to achieve significant improvement in survival. Intensive follow-up is a cost-effective strategy, with measurement of CEA being one of the least expensive tests performed, estimated to be approximately half the cost of a chest X-ray and less than one-tenth the cost of a colonoscopy in the USA.

A number of expert panels have issued broadly similar recommendations stating that CEA should be measured in patients with Dukes stage B or C CRC who may be eligible for further intervention (e.g. resection of liver metastases or systemic treatment) if their disease recurs. It should be measured at baseline and then every 2–3 months for at least three years after diagnosis and approximately every six months for a further five years. Measurement of CEA following resection of solitary liver metastases provides helpful information, with CEA falling rapidly following successful resection. Unfortunately, while CEA is a relatively sensitive marker for liver metastases, this is not the case for lung metastases.

Increasing CEA concentrations in an asymptomatic patient pose a difficult dilemma. The National Comprehensive Cancer Network (NCCN) recommends that if CEA is increasing in the presence of normal imaging, computed tomography (CT) scans should be repeated every three months until disease is detected or

CEA concentrations stabilize or decrease. Exploratory surgery based solely on an increase in CEA is not recommended by the NCCN; neither is the use of CEA-radiolabelled scintigraphy in patients with negative imaging. It is important to note that low serum CEA results do not necessarily exclude progression and other confirmatory tests (e.g. CT scan, X-ray, colonoscopy or possibly CA19-9 measurement) may be required if this is clinically suspected.

Monitoring of advanced disease

Patients with metastatic CRC are likely to receive a fluoropyrimidine (e.g. fluorouracil) in combination with oxaliplatin and/or irinotecan. Biological treatment with therapeutic antibodies (e.g. bevacizumab, cetuximab and panitumumab) is also increasingly used. While radiology remains the 'gold standard' method for evaluating response, studies have demonstrated good concordance between radiological response and CEA response in >90% of patients being treated for isolated CRC liver metastasis, leading to the conclusion that CEA is as accurate as CT imaging for assessing response to chemotherapy. Caution in interpretation is, however, required, particularly in the first few weeks following chemotherapy, as transient rises in CEA may occur in 10–15% of patients. Interestingly, these increases appear to be associated with a favourable outcome and are thought to be due to necrosis and/or apoptosis caused by the cytotoxic therapy. A confirmed increase during treatment may lead to a change of treatment or withdrawal of ineffective therapy.

Cautions and caveats

It is essential that, whenever possible, CEA concentrations in individual patients are monitored using the same method, since method-related differences in recognition of CEA mean that results may differ sufficiently to influence the interpretation of serial measurements. As for all serum tumour markers, when a change of method is unavoidable, the laboratory should establish a new baseline value for each patient.

An increase of >30% from the previous value of CEA is considered to be significant and should be confirmed by a second sample taken within a month. Smaller increases (e.g. 15–20%) demonstrated over at least three consecutive assays at minimum intervals of two weeks, may also prompt intervention. Patients with symptoms of recurrence require additional investigation such as CT scanning. CEA results within the reference interval do not exclude recurrence. Conversely, increasing CEA concentrations following curative surgery may reflect factors that are not associated with the primary tumour (e.g. lung nodules, liver lesions, ovarian masses and mediastinal lymphadenopathy).

Genetic pre-screening for hereditary non-polyposis colon cancer

Considerable progress has been made in understanding genetic influences on the development of CRC, and genetic testing is likely to become increasingly important.

Pre-screening for HNPCC, which accounts for approximately 3% of all CRC patients and which predisposes to development of other cancers, is currently performed by testing for microsatellite instability (MSI), a surrogate marker for DNA mismatch repair (dMMR) gene dysfunction, which is present in >90% of affected patients. Further testing for mutations in the *BRAF* gene (proto-oncogene B-Raf) may be desirable in some cases, as it may determine sensitivity to specific anti-tumour drugs (see p. 837). Studies have suggested that cancer rates and mortality are both decreased by close surveillance of patients with HNPCC, but there is still much debate about whether routine measurement of MSI/dMMR is desirable in all patients with CRC. Reports also suggest that MSI/dMMR status may have prognostic relevance in CRC, with the presence of MSI or defective MMR activity associated with favourable outcomes.

K-RAS mutation detection

Colorectal cancer patients with specific activating mutations in either codon 12 or 13 of *K-RAS* rarely respond to treatment with anti-epidermal growth factor receptor (EGFR) antibodies, which include cetuximab and panitumumab, both of which bind to the extracellular domain of EGFR. Expert groups have therefore recommended that patients should be tested for *K-RAS* mutations before treatment, to identify individuals unlikely to benefit from these relatively expensive drugs.

Gastric cancer

The second most common gastrointestinal cancer worldwide, gastric cancer, is frequently diagnosed only when it is at an advanced stage. Plasma concentrations of CEA and CA19-9 are raised in 20–50% of patients with advanced disease and AFP in 20–25%, but these markers are increased in <20% of patients with early stage disease. None of these has the sensitivity or specificity required for screening or diagnosis of gastric cancer, although some studies suggest that post-treatment monitoring of patients using serum CEA or CA19-9 may provide early detection of recurrence in a proportion of patients. Further data are required, together with evidence that early detection improves clinical outcome, before monitoring with either can be recommended.

Results of a recent phase III trial suggest that trastuzumab (Herceptin[®]), in combination with chemotherapy, should be implemented as standard treatment for the 15–20% of patients with HER-2-positive advanced gastric cancers. Testing for HER-2 to identify gastric cancer patients who over-express HER-2 is likely to become increasingly important, particularly as there are newer forms of anti-HER-2 therapy (e.g. lapatinib, pertuzumab and trastuzumab-TM1), which may prove more efficacious.

Gastrointestinal stromal tumours (GIST)

Gastrointestinal stromal tumours (GIST) are rare tumours that occur in the stomach, small bowel, large bowel, oesophagus or omentum. At a molecular level,

they are characterized by the presence of KIT protein (also known as CD117 antigen), whose measurement in serum is recommended by a number of expert panels. The mutational status of the *KIT* proto-oncogene is also important in predicting which patients will benefit from imatinib, a tyrosine kinase inhibitor, whose availability has dramatically improved the treatment of patients with GIST.

Germ cell tumours

Germ cell tumours are more common in males than females and can be benign or malignant. Although most frequent in young adults, they can occur at any age. They may be seminomatous germ cell tumours, non-seminomatous germ cell tumours (NSGCT) or combined tumours (tumours with both seminomatous and non-seminomatous elements). Germ cell tumours often originate in the gonads, but they also occur elsewhere, particularly in the mediastinum, retroperitoneum or pineal gland (i.e. along the 'midline'). Testicular germ cell tumours are increasing in incidence and ~1500 new cases are diagnosed each year in the UK. Plasma concentrations of AFP and/or hCG are elevated in 80–85% of men with NSGCT, whereas <25% of those with seminomas have raised hCG and none have raised AFP. Methods used in oncology for the measurement of hCG should recognize both intact hCG and its free β -subunit ('total β hCG' assay) since a significant proportion of hCG in patients with seminomas may be present as the free subunit rather than the intact molecule. Placental alkaline phosphatase (PLAP) is another promising tumour marker in seminoma, but lack of commercially available assays and the significantly increased concentrations of PLAP observed in smokers have thus far limited its application.

Although germ cell tumours are usually aggressive neoplasms, they are highly sensitive to treatment with surgery and, when appropriate, chemotherapy and/or radiotherapy. The anticipated cure rate is >90%, although patients presenting with advanced disease have a lower five-year survival of 50–60%.

There are well-accepted clinical guidelines relating to the management of patients with germ cell tumours, for which measurements of AFP, hCG and LDH are an integral part. In the UK, patients are referred to a regional centre serving a population of two to four million for further assessment and chemotherapy. All patients should be discussed at multidisciplinary team meetings attended by a clinical biochemist. Tumour marker results should be reviewed along with histological immunostaining and radiological results and any inconsistencies noted when treatment decisions are considered.

Screening

The relatively low sensitivity and specificity of AFP, hCG and LDH for germ cell tumours, together with the low prevalence of these cancers in the general population, means that tumour markers cannot be used for screening.

Diagnosis

The possibility of a germ cell tumour should be considered in any patient with a poorly defined epithelial malignancy, particularly young individuals with mid-line masses. Plasma concentrations of AFP and hCG should be measured in any male with a suspicious lump in the testis and in any patient under 50 years of age with a malignancy of unknown origin. Young patients with germ cell tumours and very high serum hCG concentrations may present with thyrotoxicosis as hCG shares structural similarities with TSH. Patients with markedly elevated concentrations of AFP and/or hCG and clinical findings consistent with a germ cell tumour (e.g. testicular lump, lung metastases, abdominal mass) should be discussed urgently with the consultant responsible for managing germ cell tumours at the regional cancer centre. Immediate referral for chemotherapy may be appropriate if surgery is likely to be required later to remove residual tumour.

Distinguishing teratomas from seminomas is essential, as their treatments differ. Production of AFP is associated with yolk sac elements in germ cell tumours and therefore AFP is not found in pure seminomas (which never contain yolk sac elements), while the presence of hCG is associated with syncytiotrophoblast tissue. Differentiated teratoma tissue does not produce either AFP or hCG. Marker measurements can sometimes modify histopathological diagnoses, e.g. raised AFP in a patient diagnosed with pure seminoma suggests that yolk sac elements may have been overlooked or that the AFP is not related to the tumour (e.g. may be of liver origin). Before surgery for a suspected germ cell tumour, measurement of both AFP and hCG is essential to allow the rate of fall of the markers to be monitored post treatment. Some patients with pure seminoma histologically may have a preoperative AFP of 5–30 kU/L (6–36 μ g/L) that remains stable following surgery. Provided that the AFP concentration remains unchanged before and after orchidectomy and is stable when re-checked four weeks after surgery, the AFP may be considered to be 'normal' for that individual and not related to the tumour.

Rarely, if primary or secondary (metastatic) disease of the central nervous system (CNS) is suspected, measurement of hCG in cerebrospinal fluid (CSF) may contribute both to diagnosis and to treatment monitoring. A CSF:serum hCG ratio of >2% is predictive of CNS involvement. In practice, measurement of tumour markers in the CSF has been largely superseded by radiological investigations.

It is important to remember that tumour marker concentrations within reference intervals do not exclude malignancy since up to 25% of NSGCT do not produce AFP or hCG, and only a small proportion of seminomas or dysgerminomas (the female equivalent of seminomas) produce hCG. The possibility of false positive results must also be considered and account should always be taken of clinical and radiological findings. Plasma AFP can be increased in benign liver disease and primary hepatocellular carcinoma as well as in other cancers, including gastric and oesophageal tumours. Similarly, hCG can be increased in other non-germ cell tumours (e.g. bladder, lung) and both AFP and hCG are significantly raised in pregnancy.

Prognosis

Tumour marker measurements contribute significantly to the prognostic assessment of germ cell tumours. Criteria developed by the International Germ Cell Cancer Collaborative Group (IGCCCG) classify patients with metastatic NSGCT as belonging to one of three prognostic groups (Table 42.6). The lowest tumour marker concentration reached post surgery, the primary tumour site, and the sites of metastatic disease all contribute to this prognostic classification.

Whether AFP and hCG measurements made prior to primary surgery are prognostically useful is less clear,

TABLE 42.6 International germ cell cancer collaborative group prognostic classification for metastatic testicular germ cell tumours

Non-seminoma	Seminoma
<p>Good prognosis Testis/retroperitoneal primary No non-pulmonary visceral metastases Good markers – all of: AFP <1000 µg/L hCG <5000 U/L (1000 µg/L) LDH <1.5 × ULN 56% of non-seminomas 5-year progression-free survival 89% 5-year survival 92%</p>	<p>Any primary site No non-pulmonary visceral metastases Normal AFP, any hCG, any LDH 90% of seminomas 5-year progression-free survival 82% 5-year survival 86%</p>
<p>Intermediate prognosis Testis/retroperitoneal primary No non-pulmonary visceral metastases Intermediate markers – any of: AFP ≥1000 and ≤10 000 µg/L hCG ≥5000 U/L and ≤50 000 U/L LDH ≥1.5 × ULN and ≤10 × ULN 28% of non-seminomas 5-year progression-free survival 75% 5-year survival 80%</p>	<p>Any primary site Non-pulmonary visceral metastases Normal AFP, any hCG, any LDH 10% of seminomas 5-year progression-free survival 67% 5-year survival 72%</p>
<p>Poor prognosis Mediastinal primary or Non-pulmonary visceral metastases or Poor markers – any of: AFP >10 000 µg/L hCG >50 000 U/L (10 000 µg/L) LDH >10 × ULN 16% of non-seminomas 5-year progression-free survival 41% 5-year survival 48%</p>	<p>No patients classified as poor prognosis</p>

ULN, upper limit of normal. Note that AFP is expressed in ng/mL rather than kU/L, the units in which many laboratories report AFP results. For most methods, results in ng/mL can be multiplied by 0.83 to obtain kU/L but laboratories should check that this conversion factor is valid for the assay used and ensure that their clinicians are aware of it. Tumour markers refer to plasma concentrations.

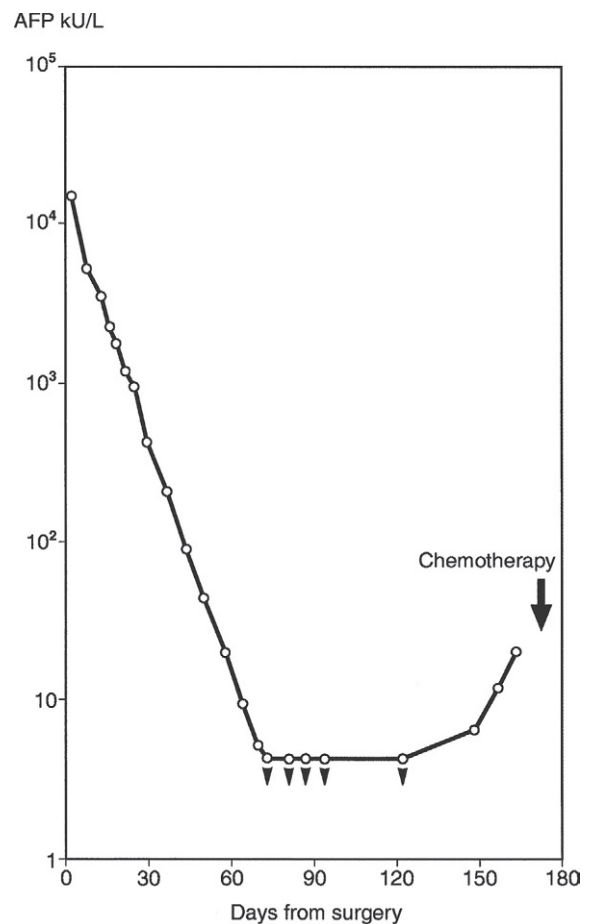
since patients with large primary tumours but without metastases may have very high AFP and hCG concentrations that return to normal following surgery.

Monitoring

Tumour markers should be measured before and after surgical excision of germ cell tumours. Ideally, if disease is limited to the testis or ovary, serum AFP and/or hCG should decrease to normal with an apparent half-life of 5–6 days for AFP and 1–2 days for hCG. This is illustrated in Figure 42.2, which shows the decrease in AFP observed in Patient A, a patient with good prognosis following surgical treatment of a malignant teratoma. The apparent half-life ($t_{1/2}$) of the tumour marker can be calculated using the equation:

$$t_{1/2} = \frac{-0.3t}{\log_{10} \frac{[M]_T}{[M]_{T_0}}}$$

where $[M]_T$ and $[M]_{T_0}$ are tumour marker concentrations at time T and T_0 , respectively and t is the difference in days between T and T_0 . As indicated in Figure 42.2, in which AFP concentration is plotted on a logarithmic



Key: ▽ indicates concentration below detection limit of assay

FIGURE 42.2 ■ Serum α-fetoprotein (AFP) concentrations in a patient treated by surgery for malignant teratoma.

scale against time on a linear scale, the apparent $t_{1/2}$ of AFP for Patient A is six days, i.e. at the upper limit of the expected half-life of AFP in normal subjects. This suggests possibly complete tumour removal. If AFP or hCG remains elevated post surgery or if metastatic disease is identified radiologically, further treatment with chemotherapy or radiation is required.

The importance of continued monitoring with tumour markers following surgery is also illustrated in Figure 42.2, where, although the measured AFP became undetectable by day 65, the concentration subsequently increased, first becoming detectable on day 151 (AFP 7 kU/L or 8.4 µg/L). Such findings should be confirmed within two weeks, after which further clinical investigations should be instituted immediately. In the absence of residual tumour or scan evidence, other causes of raised AFP and/or hCG, including the possibility of analytical interference, must be excluded, but provided this is done, Patient A would be a candidate for chemotherapy to combat progressive disease. Patients with very high marker concentrations before chemotherapy are more likely to have undetected micrometastases, as was probably the case for Patient A.

All chemotherapy regimens for germ cell tumours include platinum-based drugs, but the regimen selected will depend both on prognosis and on the results of clinical trials appropriate to that prognostic category. Following chemotherapy, AFP and/or hCG concentrations should decrease linearly when plotted as described above, but their apparent half-lives will usually be somewhat longer, i.e. up to seven days for hCG and ten days for AFP. A somewhat different pattern of AFP results is seen for Patient B (Fig. 42.3), a patient with a metastatic teratoma that did not respond to initial surgery. In some patients, AFP and/or hCG may continue to rise for up to three weeks following the start of chemotherapy, although concentrations usually start to decrease by ten days (as for Patient B). They should then fall linearly with time when

plotted as described above. However, if marker concentrations plateau, as is also illustrated in Figure 42.3 (at about day 130) this is cause for concern, suggesting either that the chemotherapy used is failing to penetrate the target tissue (which may be a large mass) or that the patient has developed drug resistance. A change in chemotherapy may then be implemented or further surgery undertaken if the tumour is considered resectable, as was the case for Patient B. Timing of surgery is important, since the outcome is likely to be most favourable if surgery is performed when the tumour marker concentration is at its nadir (i.e. the lowest concentration that can be achieved with chemotherapy).

It is estimated that with an hCG concentration of only 1 U/L, up to 100 000 tumour cells may persist. Chemotherapy should therefore be continued for some time after AFP and hCG both become undetectable, so that as many tumour cells as possible are eradicated. In some patients, results may be difficult to interpret, since chemotherapy may damage the liver (e.g. causing increased serum AFP concentrations in patients with purely hCG-producing tumours). This is particularly relevant in children.

The availability of serum tumour markers for the majority of NSGCT patients has facilitated clinical trials designed to minimize treatment toxicity for patients with good prognosis disease. By considering the rate of change of tumour marker concentrations it is possible to identify subtle differences in the effects of treatment within days, much more rapidly than the weeks it may take for significant radiological changes to become apparent.

Long-term surveillance

Tumour markers should be measured regularly following treatment, according to defined clinical protocols. Time intervals will depend on prognostic category and treatment. As is clear from Figure 42.2, any analytically significant increases in tumour markers should be reported immediately to the relevant clinical team.

Seminoma patients at low risk of recurrence who have remained well for five years can be discharged from follow-up, as can those with NSGCT who have remained well for ten years. Separate monitoring of cardiovascular risk in these patients may, however, be desirable as there is increasing evidence of an excess of cardiovascular events following chemotherapy with platinum-based drugs.

Gestational trophoblastic neoplasia

Gestational trophoblastic neoplasia (GTN) form a group of several diseases associated with pregnancy, usually involving abnormal growth of cells within the uterus. Rare and once fatal, these now highly curable tumours develop in the trophoblast cells that surround the embryo immediately after conception. Gestational trophoblastic neoplasia may develop after a molar pregnancy, a non-molar pregnancy or a live birth and should be considered in any woman developing acute respiratory or neurological symptoms or persistent abnormal vaginal bleeding after any pregnancy. In the UK, the Royal College of Obstetrics and Gynaecology has developed guidelines for management of GTN (available at: www.rcog.org.uk).

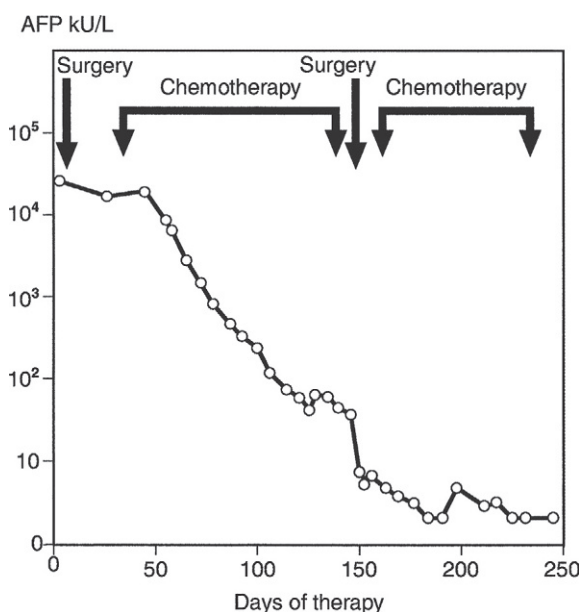


FIGURE 42.3 ■ Serum α -fetoprotein (AFP) concentrations in a patient with metastatic teratoma that was initially unresponsive to surgery.

Since measurement of hCG and related molecules is fundamental to the successful management of these diseases, it is useful briefly to review their characteristics.

Hydatidiform moles

The most common type of GTN is hydatidiform mole (also termed molar 'pregnancy', although a normal baby cannot be produced), which is not cancerous. The moles are villi that have become swollen with fluid and then grow in clusters resembling bunches of grapes. Hydatidiform moles may develop when a sperm fertilizes an 'empty' egg containing no nucleus or DNA ('complete' mole) or when two sperm fertilize a normal egg ('partial' mole). There is no fetal tissue present in complete moles, as all the genetic material comes from the father's sperm. Relatively few patients with partial moles need further treatment after initial surgery and these moles rarely become malignant. Up to 20% of patients with complete moles will need further surgery or chemotherapy; a small percentage develop into choriocarcinoma, a malignant form of GTN.

Invasive moles

Invasive moles develop in about 20% of women who have had a complete mole removed by curettage of the lining of the uterus. Invasive moles penetrate the muscular wall of the uterus (the myometrium) sometimes leading to heavy bleeding. In about 15% of patients, the tumour metastasizes to other sites, most frequently the lungs. The risk of developing invasive mole is increased if more than four months elapse between cessation of periods and treatment, if the uterus has become very large, if the woman is over 39 years old or if she has had a previous GTN.

Choriocarcinoma

Choriocarcinoma is a malignant form of GTN that most often develops from a complete hydatidiform mole but can occur after normal pregnancy or after early fetal loss in pregnancy. Choriocarcinomas are more likely to metastasize to distant organs than invasive moles, but treatment with chemotherapy is highly effective.

Placental site trophoblastic tumours

These are rare forms of GTN that develop where the placenta remains attached to the uterus, often after a normal pregnancy or abortion. They do not usually metastasize but are not sensitive to chemotherapy and must be completely removed surgically.

Screening

All patients in the UK who have had a previous molar pregnancy are automatically registered with the National Hydatidiform Mole Registry. Screening for choriocarcinoma using hCG in this highly selected group of women provides the single best example of a successful screening programme using a tumour marker, since ~8% of these women will develop the disease (i.e. the prevalence in the population studied is high) and early detection

and treatment improves outcome. The logistics are also convenient as hCG, which has approximately 99% sensitivity and specificity for choriocarcinoma, can be measured reliably in urine specimens that are sent by post by the patient to laboratories in specialist referral centres, thereby facilitating early detection of disease and prompt treatment. These laboratories use broad-spectrum hCG methods that are designed to detect hCG and its major isoforms.

Diagnosis

Increasing use of ultrasound in early pregnancy has probably led to earlier diagnosis of molar pregnancy, but measurements of hCG may also contribute as concentrations may be higher than in normal pregnancy.

Prognosis

Prognosis is assessed according to the International Federation of Gynaecology and Oncology staging for GTN, the hCG concentration prior to treatment being one of a number of factors scored. Low risk patients receive different treatment from those classified as high risk.

Monitoring

Women with persistent GTN are treated with appropriate chemotherapy at specialist centres. About 15% of patients with complete moles require chemotherapy and 0.5% after partial moles. Treatment is continued until the hCG concentration has returned to normal and then for a further six weeks. Women should be advised not to attempt to conceive until the hCG concentration has been normal for six months. After any further pregnancy, hCG should again be measured to exclude disease recurrence.

Hepatocellular carcinoma (primary liver cancer)

Hepatocellular carcinoma (HCC) is rare in the developed world but is common in China, South-east Asia and sub-Saharan Africa, and is the fifth most common cause of cancer death worldwide. In most parts of Africa and Asia, infection with hepatitis B virus is a major causative factor, as is ingestion of the fungal toxin, aflatoxin B₁, from contaminated food. The higher incidence observed in Europe during the last decade probably reflects the increased frequency of hepatitis C infection and alcoholic liver cirrhosis, both of which are strongly associated with development of HCC.

Although many potential tissue and serum tumour markers are increased in HCC, AFP is at present the most clinically useful. Normally produced during gestation by the fetal liver and yolk sac, AFP is increased in the maternal circulation during pregnancy and is also markedly elevated in newborns, with concentrations declining over the first year of life. An oncofetal antigen, AFP appears inappropriately in adults with malignancy, most frequently in hepatocellular carcinoma and germ cell cancers, and also in some benign conditions, particularly those associated with liver damage and/or regeneration. Circulating

AFP concentrations range from within the reference interval to as high as 8.3×10^6 kU/L (10×10^6 µg/L), but in the UK, up to 50% of patients with HCC may have normal concentrations of AFP.

Screening of high-risk groups

Screening by six-monthly measurement of serum AFP together with abdominal ultrasound, is now recommended for early detection of HCC in high-risk populations. There is increasing evidence that such screening (when compared with no surveillance) detects HCC of smaller size and enables a greater proportion of patients to be cured, thereby leading to improved long-term survival and cost-savings. In the UK, it is suggested that such screening be restricted to patients with liver cirrhosis secondary to hepatitis B or C or genetic haemochromatosis, and to men with primary biliary cirrhosis (PBC) and alcoholic cirrhosis (if abstinent). (Available data suggest that women with PBC or alcoholic cirrhosis have a lower risk of developing HCC.) Validation of optimal follow-up protocols, when AFP is raised or when suspicious nodules are detected, is now in progress. Sequential measurement of serum AFP provides useful information and is undergoing validation for routine clinical practice. An elevated AFP concentration detected by a single measurement may be transient (e.g. arising from an inflammatory flare of underlying viral hepatitis). Elevated but stable concentrations decrease the likelihood that HCC is the causative agent, while a steadily rising AFP concentration should always be rigorously investigated.

Diagnosis

Plasma concentrations of AFP up to 1245 kU/L (1500 µg/L) may occasionally be associated with benign conditions, with 20–40% of adult patients with hepatitis or liver cirrhosis having AFP >8.3 kU/L (10 µg/L). An AFP result within the reference interval does not exclude a diagnosis of HCC. Rapid increases (a doubling time of <5 days) are suggestive of acute liver damage rather than malignancy. As with screening, a steady rise in AFP is suggestive of HCC, while stable or decreasing results make HCC less likely. Histopathology of appropriate biopsy material is essential for definitive diagnosis of HCC since AFP may be raised in malignancies other than HCC, including germ cell tumours and stomach, biliary tract and pancreatic cancers.

Prognosis

Large multivariate analyses confirm that raised AFP concentrations predict poor prognosis when compared with AFP-negative patients, some studies indicating that patients with larger tumours tend to have higher AFP concentrations. Together with tumour size and extent, AFP appears to be an independent predictor of survival, patients with serum AFP >8300 kU/L (10 000 µg/L) doing less well than those with AFP <166 kU/L (200 µg/L). α-Fetoprotein doubling time may also be an important prognostic factor.

Monitoring

Following complete surgical removal of HCC, AFP concentrations typically decrease with a half-life of 3.5–4 days, incomplete resection being associated with a longer half-life and poorer outcome. α-Fetoprotein concentrations that fail to return to normal suggest incomplete resection or severe liver damage.

Serum AFP measurements can also be used to monitor HCC patients following treatment with chemotherapy or radiotherapy. Decreases in tumour markers may in some cases reflect tumour regression more accurately than CT scans, since interpretation of the latter may be complicated by residual fibrosis and other factors. It is important to note that recurrence is possible even when AFP is stable or within reference limits, presumably from micrometastases too small to produce measurable AFP concentrations in serum, and that recurrent tumour may not secrete AFP even if the original tumour did so.

Lung cancer

Lung cancer is the cancer with both the highest incidence and the highest number of deaths in the world. There are two major histological types: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). Non-small cell lung cancer, which accounts for 75–85% of lung cancers, consists of several subtypes, predominantly squamous cell carcinoma, adenocarcinoma and large cell carcinoma. Surgery is the only curative treatment for NSCLC. Small cell lung cancer accounts for 15–25% of lung cancers, often has neuroendocrine elements and is primarily treated with chemotherapy and/or radiotherapy. Small cell lung cancer is an aggressive tumour characterized by a short doubling time and early development of metastatic disease. Many lung cancers are mixed tumours containing both small cell and non-small cell components.

A large number of tumour markers have been tested in lung cancer, different markers being required to identify the response to treatment of the different cell types present. Some of the more promising markers and their associated cell types are listed in Table 42.7.

TABLE 42.7 Tumour markers of potential utility in different types of lung cancer

Tumour marker	Type of cancer	Cell type
Cyfra 21-1	NSCLC	Squamous cell
Carcinoembryonic antigen (CEA)	NSCLC	Adenocarcinoma Large cell
Squamous cell carcinoma antigen (SCC)	NSCLC	Squamous cell
Neuron specific enolase (NSE)	SCLC	Small cell (neuroendocrine)
Progastrin releasing peptide (ProGRP)	SCLC	Small cell (neuroendocrine)

NSCLC, non-small cell lung cancer; SCLC, small cell lung cancer.

Screening

No markers are suitable, either singly or in combination, for screening either in the general population or in specific high-risk groups, such as smokers.

Differential diagnosis

The pattern of elevated serum concentrations of CEA, Cyfra 21-1, NSE, SCC and ProGRP may suggest which histological subtype of lung cancer is present but further work is required before measurement of tumour markers in this situation becomes standard practice. Where inoperable lung cancer is suspected but no histology is available, raised serum NSE and especially ProGRP concentrations are highly suggestive of small cell lung cancer, while raised serum SCC concentration is suggestive of squamous cell cancer.

Prognosis

Studies have suggested that Cyfra 21-1, CEA, NSE and LDH (lactate dehydrogenase) may all provide prognostic information in NSCLC, whereas only NSE and LDH act as prognostic indicators in SCLC.

Monitoring

Although tumour markers can be used in individual lung cancer patients to monitor their response to therapy (using NSE and/or ProGRP for SCLC, and CEA and/or Cyfra 21-1 for NSCLC), in view of the limited range of treatment options currently available, this is controversial. However, serial determinations of the appropriate marker following surgery may help to assess the completeness of tumour removal and provide early indication of recurrence.

In patients receiving systemic treatment, measurements of tumour markers may assist in assessing response to therapy and to document progressive disease, although reliable criteria for 'biochemical progression' have yet to be developed. There is currently no evidence that monitoring with tumour markers improves patient outcome.

Epidermal growth factor receptor and K-RAS mutation analysis

A number of expert panels recommend determining *EGFR* mutation status in patients with advanced NSCLC prior to administration of the tyrosine kinase inhibitors (TKI) gefitinib or erlotinib. Patients lacking the mutation are unlikely to benefit, while 65–70% of those with specific activating mutations in the *EGFR* gene respond to these therapies. Similarly, there is a significant correlation between the presence of *K-RAS* mutations and an absence of response to the TKIs. Combining the tumour mutation status of *K-RAS* with that of *EGFR* can therefore be helpful both in identifying patients with advanced NSCLC who are likely to respond to treatment with TKIs and the likelihood of resistance to specific TKIs.

Melanoma

Melanoma is a malignant tumour of melanocytes (cells that are derived from the neural crest), with a rising incidence worldwide. Most melanomas occur in the skin, but they may also develop at mucosal surfaces or other sites to which neural crest cells migrate. Early stage disease can be cured by surgery but the prognosis associated with metastasis to distant sites is poor.

Immunohistochemical staining with antibodies to S100 is the method of choice for diagnosing malignant melanoma in pathological specimens, but further studies are required to define the role of S100 measurement in serum. Measurement of S100B lacks sensitivity in early disease. Rising concentrations of S100B are specific and sensitive for tumour progression in patients with advanced disease, but measurement is only appropriate if further treatment options are available. Measurement of a tumour-associated glycoprotein antigen (TA90-IC) shows promise as a prognostic indicator and for monitoring patients but requires further evaluation. Lactate dehydrogenase, although a very non-specific marker, can be useful for monitoring patients with melanoma and may have prognostic value in patients with advanced disease.

BRAF mutation analysis

About 40–50% of patients with disseminated disease, but without brain metastases, may respond to treatment with the *BRAF* gene inhibitor vemurafenib, provided they have a V600 mutated *BRAF* gene. *BRAF* gene status therefore needs to be confirmed prior to treatment.

Neonatal and paediatric tumours

Cancer is the second leading cause of death in children <15 years old, but major advances in treatment mean that >70% of children diagnosed with cancer are now cured. Most childhood solid tumours are of mesenchymal or embryonal origin. Tumour markers can contribute significantly to the management of childhood neuroblastomas, malignant hepatic tumours and germ cell tumours. While the general principles of marker use in these cancers are the same as for adult malignancies, some additional points should be noted.

When interpreting AFP, it is important to remember that plasma AFP is markedly raised at birth and then declines steadily to adult concentrations by 6–12 months. α -Fetoprotein is higher in children born prematurely and may remain elevated for longer in children with delayed development. Appropriate gestational and age-related reference intervals therefore need to be used for infants. Serial concentrations are often more useful than isolated results. This is particularly relevant in neonates, in whom acute hepatocellular damage may result in marked increases in AFP concentrations. As with adults, other possible causes of raised AFP need to be considered, including hereditary tyrosinaemia and ataxia telangiectasia. The very high concentrations of tumour markers that may be seen in some childhood cancers mean that particular care must be taken to minimize the risk of the high-dose hook effect (Box 42.1). Since AFP and hCG requests

for young children are relatively infrequent, it would be eminently feasible, and highly desirable, to assay all such samples at more than one dilution.

Germ cell tumours in childhood

As in adults, AFP and hCG are often elevated at the time of diagnosis and their measurement is mandatory. Yolk sac tumours are the most common pure malignant germ cell tumours in children. Seminomas rarely occur in infants or young boys, but dysgerminomas are the most common pure malignant germ cell tumour occurring in the ovary and central nervous system in girls, who may present with precocious puberty.

Hepatoblastoma

Hepatoblastoma and hepatocellular carcinoma (HCC) are the most frequent malignant hepatic tumours of childhood. More than 80% of hepatoblastomas, which are embryonal tumours, are diagnosed in children aged <3 years, with 45% of patients diagnosed during the first year of life. Most patients (98%) have a raised AFP at presentation, often to extremely high concentrations (e.g. 10^6 kU/L or 1.21×10^6 µg/L), which can assist in diagnosis. Thereafter, serum AFP can be used to monitor therapy and follow-up. Children with hepatoblastomas that secrete hCG may develop isosexual precocious puberty. Complete surgical resection is the treatment of choice, with chemotherapy also playing an important role. Overall survival rates are >60%, or >80% if complete resection is achieved.

Hepatoblastomas must be differentiated from HCCs, 50% of which also produce AFP. Positive hepatitis B serology is present in some children with HCC, with other laboratory abnormalities including anaemia and hyperbilirubinaemia. Complete surgical resection is the treatment of choice for HCC. Aggressive chemotherapy has not significantly improved outcome and most children with HCC die within 12 months of diagnosis.

Neuroblastoma

Neuroblastoma is a malignant embryonal tumour that accounts for 8–10% of all childhood cancers, with 80% of cases occurring before the age of four years. The clinical behaviour is varied, some tumours undergoing spontaneous regression and others exhibiting extremely malignant behaviour. Treatment includes surgery, chemotherapy and radiotherapy. Urinary catecholamine metabolites are increased in >90% of patients and are helpful to confirm the diagnosis and monitor progress (see Chapter 38). Significant elevations of serum NSE, LDH and/or ferritin tend to be associated with a poorer outcome.

Ovarian cancer

Ovarian cancer is the fourth most common cause of death from cancer in women in the UK. Early diagnosis is key to successful treatment, but the absence of symptoms in early stage disease means that many ovarian cancers present late. Standard treatment for ovarian cancer is surgical, involving

bilateral oophorectomy and pelvic clearance, usually followed by chemotherapy. New chemotherapeutic agents have significantly improved the five-year survival rate.

About 15% of malignant ovarian tumours are germ cell tumours, for which AFP, hCG and LDH are the markers of choice as previously described, or sex cord stromal tumours, two-thirds of which are granulosa cell tumours. Inhibin is the marker of choice for granulosa cell tumours of the ovary. An inhibin method, which detects all forms of inhibin, including A, B and pro-αC, is required.

However, most malignant ovarian tumours (80–85%) are surface epithelial carcinomas. These occur in five histologically distinct subtypes: serous, mucinous, endometrioid, clear cell and transitional. They exhibit different clinical behaviours, tumorigenesis and pattern of gene expression, which should be taken into account when evaluating the clinical utility of tumour markers in ovarian cancer. The most widely used tumour marker for epithelial ovarian cancer is CA125. Most sensitive in serous adenocarcinomas, it is also used in other histological types of epithelial cancer despite its poorer sensitivity.

Interpretation of CA125 results is particularly challenging, and it is essential to be aware of the many benign conditions in which the marker can be significantly elevated, sometimes transiently (Table 42.1). These conditions include during menstruation and the first two trimesters of pregnancy. Serum CA125 may be raised in any patient with ascites (occasionally to >5000 kU/L) or pleural effusion (usually 200–300 kU/L) and is also raised in patients with a wide variety of tumours, especially adenocarcinomas. Method-related differences in results can also be problematic and some methods are more vulnerable to interference, such as from heterophilic antibodies, as illustrated in Table 42.8.

Screening

In view of the absence of early symptoms, a reliable screening test for ovarian carcinoma would be highly desirable. Variations in results in premenopausal women (including increases during menstruation) mean that

TABLE 42.8 Misleading results for CA125 owing to presumed antibody interference in an automated immunoassay method

Day	CA125 (kU/L)		Treatment
	Method 1	Method 2	
		Undiluted specimen	
0	–	–	Laparotomy
20	130	346	578
41	39	420	566
49	–	–	–
76	22	>500	708
104	19	>500	1040
132	21	302	417
160	16	166	267
202	16	274	336
278	16	35	47
			Follow-up

single CA125 measurements lack the sensitivity and specificity essential in a screening setting. However, use of CA125 for screening in postmenopausal women is the subject of several trials. In a large controlled trial (the UK Collaborative Trial of Ovarian Cancer Screening), patients have been randomized into three groups: an unscreened control arm; a group screened annually with CA125, with repeat CA125 and ultrasound follow-up if positive, and a group screened annually with ultrasound, with repeat ultrasound in 6–8 weeks if positive. Until this trial has reported in 2015, screening of the general population cannot be recommended.

There is, as yet, no evidence that using CA125 to screen women at high risk of ovarian cancer (e.g. those with a strong family history) is effective. However, a US National Cancer Institute Panel has recommended annual CA125 determinations, in addition to pelvic and ultrasound examinations, in women with a history of hereditary ovarian cancer, who have an estimated lifetime risk of 40%. Such investigations should always be carried out in specialist units.

Diagnosis

Ultrasound assessment is essential in women with suspected ovarian cancer. This is likely to identify a pelvic mass and may suggest the presence of metastatic disease. If no obvious source is identified, determining whether the pelvic mass is likely to be malignant can be problematic but is important, as it will influence plans for surgery. Prognosis in ovarian cancer correlates strongly with the extent of surgical clearance of malignant tissue.

Cut-off concentrations (e.g. CA125 >95 or >65 kU/L) have been used with some success in postmenopausal women to distinguish between malignant and benign pelvic masses. However, a risk of malignancy index (RMI) scoring system, incorporating CA125, ultrasound findings and menopausal status, is more widely used (Table 42.9). The two scoring systems (RMI 1 and RMI 2) are similar but RMI 2 gives greater weight to ultrasound findings and menopausal status than does RMI 1. A positive predictive value for malignancy of

about 80% is obtained using the RMI 2 scoring system and a cut-off value of 200.

It is clear from Table 42.9 that whichever score is used, the result depends critically on the numerical CA125 value and therefore also on the reliability of the assay, particularly in the range of 10–25 kU/L. If there is a change in method the cut-off value used may need to be reviewed.

In the UK, NICE has published clinical guidelines on the recognition and initial management of ovarian cancer with the aim of increasing awareness of the disease among general practitioners and decreasing the frequency of late diagnosis, which is thought to account for the lower survival rates observed in the UK and Ireland compared with other European countries. The NICE guidelines state that general practitioners should measure serum CA125 concentrations in women presenting with persistent and continuous symptoms (>12 times a month) suggestive of ovarian cancer (e.g. abdominal pain, early satiety and loss of appetite or pelvic pain). If CA125 concentration is ≥ 35 kU/L, an ultrasound scan of the abdomen and pelvis should be arranged to enable calculation of the RMI. Women with an RMI score ≥ 250 should be referred to a specialist multidisciplinary team. Women with CA125 <35 kU/L should be reviewed at six weeks if symptoms persist.

Prognosis

During primary treatment, CA125 concentrations both before and after surgery may be of prognostic significance. Patients with a preoperative CA125 concentration >65 kU/L have a poorer five-year survival than patients with lower CA125 concentrations. A concentration of <250 kU/L before chemotherapy, a fall of CA125 of greater than seven-fold during the first month of chemotherapy, an apparent CA125 half-life of <20 days during chemotherapy and a CA125 concentration of <35 kU/L before the third course of chemotherapy are all thought to indicate a good prognosis.

Detection of residual disease

Measurement of serum CA125 following surgery is only relevant in the small number of patients with disease limited to the ovary (Stage I) who would not automatically receive further treatment. It is essential that these patients have CA125 measurements at least once a week initially so that the rate of fall can be checked. The apparent half-life of CA125 in patients with no residual disease is five days.

Monitoring

Serum CA125 concentrations correlate well with the response to treatment in ~90% of patients. A rising or stable CA125 at the beginning of the second or third course of chemotherapy, together with an apparently poor clinical response, may be used as an indication either to change the treatment or to institute appropriate palliative therapy. While an increased CA125 concentration (>35 kU/L) at the end of treatment is always associated

TABLE 42.9 The risk of malignancy index (RMI) scoring system for ovarian cancer

Feature	RMI 1 score	RMI 2 score
Ultrasound features:		
multilocular cyst solid areas	0=no abnormalities 1=one abnormality 3=two or more abnormalities	0=no abnormalities
bilateral lesions		1=one abnormality
ascites		4=two or more abnormalities
intra-abdominal metastases		
Premenopausal status	1	1
Postmenopausal status	3	4
Plasma CA125 (kU/L) concentration	Pre-treatment value	Pre-treatment value

RMI score = (ultrasound score) × (menopausal score) × (CA125 concentration in kU/L).

with disease, the reverse is not true, as a low CA125 concentration does not exclude active disease.

Long-term surveillance

A confirmed rise in CA125 to more than twice the upper reference limit accurately predicts relapse with a sensitivity of 86% and a positive predictive value of 95%. Whether early treatment for relapse in patients who are asymptomatic improves outcome, is still unclear. Results of a Medical Research Council and European Organisation for Research and Treatment of Cancer randomized control trial (MRC OV05 and EORTC 55955) addressing this question suggests that instituting early chemotherapy based on a rising CA125 does not improve either survival or quality of life, and that it is appropriate to delay treatment until signs and symptoms of recurrence develop. Whether this would also be the case with the improved treatment regimens now available is not clear. However, the recommendation from the trial investigators is that it is appropriate to offer women with ovarian cancer two informed choices for follow-up, i.e. no routine CA125 measurements but rapid access to CA125 testing if there are symptoms or signs of relapse, or, alternatively, regular CA125 measurements.

Pancreatic cancer

In pancreatic ductal adenocarcinomas (i.e. non-endocrine tumours of pancreas), CA19-9 is the only tumour marker for which there is sufficient evidence to support its clinical use. It can be used as an adjunct to diagnosis in association with appropriate imaging but it is not suitable for screening even in high-risk populations. Measurements can provide independent prognostic information with regard to resectability and survival but should only be used in conjunction with other clinical information. The American Society of Clinical Oncology recommends serial measurements of CA19-9 every 1–3 months for pancreatic cancer patients with locally advanced or metastatic disease who are receiving active therapy. Increases in serial CA19-9 measurements suggest progressive disease, but confirmation by other means should be sought. It is important to note that there are significant method-related differences in CA19-9 results and particular care should be taken if changing methods. Many non-malignant conditions, including any that cause cholestasis, can cause elevated plasma concentrations of CA19-9 (see Table 42.2).

Prostate cancer

Prostate cancer is one of the most common malignancies in men, autopsy data indicating the presence of histologically apparent cancer in the prostates of ~42% of men aged >50 years who have died from other causes. Recent increases in the apparent incidence of prostate cancer almost certainly reflect widespread measurement of serum PSA, elevations of which can identify disease long before it is symptomatic. Since many men have indolent prostatic cancers that pose little threat to their life or health, there is increasing concern that the use of PSA measurements,

which cannot differentiate slow-growing cancers from aggressive cancers that require treatment, is causing over-diagnosis and over-treatment of some men. A further difficulty is that, while PSA is essentially organ-specific, it is not cancer-specific, and men with non-malignant conditions, such as benign prostatic hypertrophy, may also have raised PSA concentrations. This lack of specificity complicates the interpretation of results, as well as having major implications for use of PSA in screening. Many uncertainties also remain concerning optimal treatment of early stage disease, even when clinically localized to the prostate.

Screening and diagnosis

Screening men for prostate cancer using PSA is controversial, as is evident from the disparate recommendations, both for and against screening, that have been made by different professional organizations. Consensus is unlikely to be achieved until publication of the final conclusions of a number of major national and international prospective screening trials using PSA (with or without digital rectal examination). Results from the European Randomized Study of Screening for Prostate Cancer, which involved 182 160 men age 50–74 years at entry, showed that, at 11 years of follow-up, PSA-based screening significantly reduced mortality from prostate cancer but did not affect all-cause mortality. To prevent one death from prostate cancer at 11 years of follow-up, 1055 men would need to be invited for screening and 37 cancers would need to be detected. In contrast, after 13 years of follow-up, results from the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial in the USA showed no evidence of a mortality benefit for organized annual screening as compared with opportunistic screening. Various explanations for these apparently contradictory results have been put forward.

However, whatever the final outcome of the ongoing trials, men will continue to request *ad hoc* screening for prostate cancer, whether or not an official programme is established. There is therefore general agreement that men should be informed of the benefits and limitations of PSA measurement before deciding whether to have the test performed, particularly if asymptomatic. In the UK, the NHS Cancer Screening Committee's Prostate Cancer Risk Management Programme provides helpful information about the test (Informed Choice Programme, see Further reading).

Early detection is likely to be most effective in younger men (in their 50s and 60s) who are at high risk of developing prostate cancer, African American men and men with a family history of prostate cancer. It is unlikely to benefit men with a life expectancy of less than ten years. Men with a strong family history (e.g. a first-degree relative diagnosed before the age of 60) may have a 3–4-fold higher than average risk of developing prostate cancer and should be carefully investigated.

Considerable effort has been made to improve the specificity of PSA assays, for example by considering the ratio of free to total PSA (the proportion of circulating free PSA is higher in benign disease), by assessing the rate of doubling of PSA (doubling time is more rapid in

malignant than benign disease) or by implementing age-related reference intervals (PSA may increase with age as the size of the prostate gland increases). Promising data are also emerging regarding proPSA, a proenzyme molecular form of free PSA, and a few truncated proPSA isoforms, but these tests are still under evaluation and are not widely used.

While each of these measures improves the differentiation of benign and malignant prostatic disease, whatever PSA decision limit is selected and however the marker is used, a significant number of men with benign prostatic disease will still be recommended for biopsy on the basis of a high PSA result. Conversely, a significant number of men with malignancy but with PSA below the decision point will be missed. It is estimated that up to 20% of men with a clinically significant prostate cancer will have a normal serum PSA concentration, and as many as two in every three men with an elevated PSA do not have prostate cancer detectable at biopsy. Nevertheless, PSA is currently the best routinely available serum tumour marker for screening or diagnosis of prostate cancer. It has been suggested that screening might be made more effective by starting screening earlier and referring patients with screen-detected cancers for treatment at specialist centres, and be made less harmful by avoiding screening in older men and promoting active surveillance with well-organized follow-up programmes for those with disease that is of low risk of progression.

Management

Once a diagnosis of prostatic cancer has been confirmed by finding malignant cells in a biopsy specimen, the treatment selected depends on whether disease is confined to the prostate gland or has spread to other organs. Using predictive tables combining the pre-treatment PSA concentration with information on clinical stage and the immunohistochemical Gleason score (which provides a measure of the degree of differentiation in biopsy tissue and ranges from 2 in well-differentiated tumours to 10 in completely anaplastic cancers), a reasonable prediction of the stage of localized prostate cancer can be made. For example, patients with a Gleason score of ≤ 6 and a PSA of $<10 \mu\text{g/L}$ are unlikely to have bone metastases.

Such information can then be used to select the most appropriate treatment option. If the disease has not spread beyond the prostatic capsule, possibilities include radical prostatectomy (complete removal of the prostate), brachytherapy (insertion of radioactive needles to ablate the prostate) and external beam irradiation. Active monitoring, i.e. regular clinical assessment together with measurement of PSA, is as appropriate as active intervention in some patients with localized tumours. The benefit of treatment is not yet proven, except in men with localized disease who are candidates for radical prostatectomy. In these men, the absolute reduction in the risk of death after ten years is small, but the reduction in the risks of metastasis and local tumour progression is substantial.

Following radical treatment, PSA should decrease to undetectable concentrations, with measurable concentrations providing evidence of residual disease. However, undetectable PSA concentrations do not necessarily

indicate cure. Similarly, while rising PSA concentrations following radical treatment provide biochemical indication of recurrent disease, clinical symptoms may never occur or may appear many years later. The use of ultrasensitive PSA methods with low detection limits (usually $<0.005 \mu\text{g/L}$), which permit very early identification of rising PSA concentrations, therefore remains controversial and is probably most appropriate within the context of clinical trials.

Treatment options for patients with prostatic cancer that has spread beyond the prostate (most frequently to local lymph nodes and bone) include endocrine therapy (usually by androgen blockade as prostate cancer is hormone-sensitive), radiotherapy and chemotherapy. Men with serum PSA $>100 \mu\text{g/L}$ and clinical, biochemical or radiological evidence of metastatic disease may, in some cases, be treated without histopathological confirmation of the diagnosis, while in other patients, active monitoring may be appropriate.

A major clinical application of PSA is in evaluating the efficacy of therapy, sustained increases providing evidence of disease progression, usually earlier than other diagnostic procedures. An increase in PSA should always be confirmed, and intraindividual variation in PSA of up to 20–30% taken into account before any increase is considered to be clinically significant. The doubling time of PSA can give some indication of the likely time to clinical progression. A stable PSA concentration does not necessarily exclude progression if clinically suspected. In patients with advanced, androgen-independent prostate cancer, measurement of alkaline phosphatase may be helpful to detect bone involvement. As knowledge of increasing PSA concentrations may have adverse psychological consequences, it may be desirable to stop measuring PSA in some patients, particularly if effective alternative therapy is not available.

Analytical and reporting requirements

Rigorous adherence to the quality control measures described at the beginning of this chapter is essential. The NHS Prostate Risk Management Programme formally requires that PSA assays used in the Informed Choice programme be accurately calibrated in terms of the International Standard for PSA (Table 42.4) and that they recognize free and complexed PSA equivalently (i.e. are equimolar). Screening programmes place particularly high demands on analytical performance, since relatively minor differences in performance have a major influence on outcome, i.e. whether or not a prostatic biopsy is advised. Even when PSA assays are accurately calibrated and equimolar, results for individual patients obtained using different methods are not necessarily interchangeable. It is therefore strongly recommended that laboratory reports for PSA state both the name of the assay used and the appropriate clinical decision limits.

Testicular cancer

More than 90% of testicular tumours in adults are germ cell tumours: these are discussed above. Leydig and Sertoli cell tumours, which are rare, develop in the supportive

hormone-producing tissue of the testicles (the stroma). These tumours represent 4% of adult testicular tumours and 20% of childhood tumours. Tumours of the Leydig cells (which normally produce androgens) do not usually spread beyond the testicle and can be cured by surgical removal. Those that do spread are resistant to chemotherapy or radiation. Tumours of the Sertoli cells (which normally support sperm-producing cells) are similar to Leydig cell tumours and are also difficult to treat if they spread beyond the testes. Inhibin is the marker of choice for Sertoli and Leydig cell tumours and the same analytical issues apply as for granulosa cell tumours of the ovary.

Thyroid cancer

Thyroid cancer is a relatively rare tumour, accounting for approximately 2% of cancers in all age groups and 4% of cancers in those <20 years old. There are four major types: papillary, follicular and anaplastic thyroid cancer and medullary carcinoma of the thyroid (MCT). Calcitonin is an excellent marker for MCT, which can be associated with multiple endocrine neoplasia (see Chapter 41). Papillary and follicular cancers are differentiated epithelial thyroid cancers and can be monitored using thyroglobulin, a large glycoprotein synthesized by the follicular cells and stored in the colloid space. Thyroglobulin is present in benign and malignant thyroid tissue.

Screening, diagnosis and prognosis

Serum thyroglobulin measurement has no role in the screening for, or diagnosis of, thyroid cancer and is of no value as a prognostic indicator. However, immunohistochemical detection of thyroglobulin can be useful in identifying the thyroid as the site of primary tumour in individuals presenting with metastatic disease from an unknown primary.

Monitoring

Thyroglobulin has an important role in monitoring patients with follicular or papillary thyroid carcinoma after treatment with surgery and/or radioiodine. Since thyroid stimulating hormone (TSH) stimulates thyroid follicular cells to produce thyroglobulin, serum thyroglobulin concentrations will be higher in patients with a raised TSH if there is remaining thyroid tissue or in the presence of recurrent tumour. TSH should therefore always be measured at the same time as thyroglobulin.

Detectable concentrations of thyroglobulin may be appropriate in patients treated by surgery alone, provided they are stable, but not in patients who have been treated with radioiodine to ensure total ablation of any remaining thyroid tissue. In these patients, serum thyroglobulin concentrations should decrease to <2 µg/L following ablation. In an individual with suppressed TSH, a rising thyroglobulin suggests recurrent tumour.

Historically, the difference in thyroglobulin values on and off thyroxine replacement was used to assess the presence of residual or recurrent tumour. This can now be done more effectively by measuring plasma thyroglobulin

concentration after administering recombinant TSH (rhTSH).

Analytical and reporting requirements

It is important to be aware of several potential pitfalls in thyroglobulin measurement. As for other tumour markers, it is highly desirable that specimens for thyroglobulin measurement should be taken prior to any invasive procedure. A common cause of very high serum thyroglobulin results is inappropriate specimen collection following fine needle aspiration biopsy of the thyroid.

Analytically, radioimmunoassay (RIA) and immunometric methods for thyroglobulin are particularly vulnerable to interference from anti-thyroglobulin auto-antibodies (TgAb), which may be present in some patient sera. Results obtained may be falsely low or falsely high, with immunometric methods being particularly prone to giving falsely low thyroglobulin results. Thyroglobulin antibodies should be measured in the same sample as thyroglobulin, using a sensitive immunoassay method. The presence of antibodies usually invalidates the thyroglobulin result. Comparison of results with RIA results available from specialist laboratories is informative. When interference is present, RIA results may be more reliable. It is important to note that interference may be absent even in the presence of antibodies or conversely present in their absence. Appropriate cautions should be added to patient reports.

Cancers of unknown primary

Cancers of unknown primary origin account for ~3% of all new cancer diagnoses and generally have a poor outlook, with median survival of less than a year. Although it is not clear to what extent tumour marker measurements may influence the clinical management of cancers of unknown primary, for a few patients the benefit may be considerable, enabling identification of appropriate treatment. Diagnoses for which this is most likely to be the case, together with the tumour marker measurements of relevance, are gestational trophoblastic neoplasia (hCG), germ cell tumours of the testis or ovary (hCG and AFP) and prostate cancer (PSA). These are potentially curable malignancies, which, once diagnosed, can rapidly respond to appropriate treatment. The NICE guidelines for metastatic malignant disease of unknown primary origin recommend measurement of serum AFP and hCG (particularly in the presence of midline nodal disease), PSA in men, CA125 in women with peritoneal malignancy or ascites, and a myeloma screen when there are isolated or multiple lytic bone lesions.

SUMMARY

Measurements of tumour markers already make an important contribution to the management of cancer patients. Maximizing their clinical usefulness and cost-effectiveness provides an excellent opportunity for close liaison between laboratory and clinical staff. Tumour

BOX 42.2 Key points for using tumour markers in the clinic

- A negative result for a serum tumour marker, either prior to diagnosis or following treatment, does not exclude active disease.
- A pre-treatment result is essential and provides the baseline against which subsequent results can be assessed.
- Serial results are nearly always more useful than single isolated results and frequency of marker measurement should follow guidelines where these exist.
- If decisions about proceeding to surgery or chemotherapy depend on a single tumour marker result, the result should be confirmed.
- Results that are not in accord with the clinical picture should be investigated and if necessary re-checked by another method.
- Tumour markers are seldom tissue-specific and therapy may result in a transient increase in the serum concentration because of release of tumour marker from normal tissue.
- Tumour markers should always be interpreted in the context of all available information, including clinical findings, imaging investigations and other blood tests. The possible influence of other factors (e.g. abnormal renal or liver function) should be carefully considered.

marker requests should be made only when the results are likely to be of benefit; pre-analytical and analytical requirements should be met and those using tumour markers clinically should be familiar with the key points summarized in Box 42.2.

It is likely that over the next decade, new analytical techniques such as proteomic profiling and novel genetic tests will lead to the development of more targeted tumour markers. Their introduction into routine practice will provide new challenges, which clinical biochemists are well positioned to meet.

Further reading

- Cancer Research UK website. <http://www.cancerresearchuk.org/cancer-help/>; [Accessed October 2013].
Patient-oriented source of helpful information about most cancers.
- Diamandis EP, Fritsche HA, Lilja H et al. editors. *Tumour markers. Physiology, pathobiology, technology, and clinical applications.* Washington: AACR Press; 2002.
Textbook providing in-depth coverage of the general principles of tumour marker use, organ-specific tumour markers, genomic and proteomic approaches for biomarker discovery and emerging tumour markers.
- National Academy of Clinical Biochemistry website. <http://www.aacc.org/members/nacb/lmpg/pages/default.aspx#> [Accessed October 2013].
The NACB's Laboratory Medicine Practice Guidelines (LMPG) for Use of Tumour Markers in the Clinic, focus particularly on laboratory aspects of tumour marker use in 16 major malignancies, with separate sections on quality requirements and new technologies.
- National Cancer Institute (United States National Institutes of Health) website. <http://www.cancer.gov/> [Accessed October 2013].
Information about most cancers for patients and for health professionals.
- National Comprehensive Cancer Network (NCCN) website. http://www.nccn.org/professionals/physician_gls/default.asp [Accessed October 2013].
The NCCN is an alliance of 20 of the world's foremost cancer centres. Detailed Clinical Practice Guidelines for most major malignancies are available on the website.
- National Health Service (NHS) Cancer Screening Programmes website. <http://www.cancerscreening.nhs.uk> [Accessed October 2013].
Source of current information about NHS screening policy with respect to established and pilot screening programmes, including those for breast cancer (mammography), cervical cancer, (smear testing) colorectal cancer (FOB testing) and prostate cancer (Informed Choice Programme).
- National Institute of Health and Care Excellence (NICE) website: <http://www.nice.org.uk/> [Accessed October 2013].
Evidence-based guidelines on the management of a number of cancers, including breast cancer, ovarian cancer, prostate cancer and metastatic malignant disease of unknown primary origin.
- Royal College of Obstetrics and Gynaecologists. Green-top 38: the management of gestational trophoblastic disease 2010, <http://www.rcog.org.uk> [Accessed October 2013].
Evidence-based guidelines on the treatment of gestational trophoblastic neoplasia.
- Scottish Intercollegiate Guidelines Network (SIGN) website. <http://www.sign.ac.uk/guidelines/published/numlist.html> [Accessed October 2013].
Evidence-based guidelines for more than 12 cancer sites (including breast, colorectal, lung and testicular cancers), are available on the SIGN website, together with guidelines relating to other aspects of medicine.

Molecular clinical biochemistry

Roberta Goodall

CHAPTER OUTLINE

INTRODUCTION 844

GENES AND GENE EXPRESSION 844

What is a gene? 844

Gene expression 846

Mutation, the source of diversity and disease 846

Genesis of an individual: the formation of gametes 849

Genes in families and populations 850

The variable expression of genetic disease 851

THE TECHNIQUES OF GENETIC ANALYSIS 852

Detection of specific sequences in DNA 852

Detection of mutations 855

THE APPLICATIONS OF DNA ANALYSIS 860

Diagnosis of index cases 860

Prenatal diagnosis 860

Screening 861

Pharmacogenetics 861

Inherited diseases – some examples 862

Multifactorial and polygenic disease 866

Cancer genetics 867

GENE THERAPY 869

Stem cells in gene therapy 870

Gene therapy in cancer 871

CONCLUSION 871

INTRODUCTION

If we distinguish the actual combination of genes possessed by an individual, that is the genotype, from the observable activity of those genes, the phenotype, then study of inherited disease in clinical biochemistry laboratories has traditionally been concerned exclusively with analysis of phenotypes. The last two decades have witnessed a dramatic change in this situation: molecular biological techniques are now a much more common part of the repertoire of clinical biochemistry laboratories. Initially, the identification of genes responsible for inherited diseases involved heroic efforts, requiring expensive, complex and extremely time-consuming procedures plus a few strokes of luck. Once a gene has been identified, however, modern analytical techniques make detection of mutations more straightforward than before.

Although each nucleated human cell contains about two metres of deoxyribonucleic acid (DNA) (around 3 billion bases), it is the fundamental simplicity of DNA – its building blocks comprise just four nucleotides – that favours its automated analysis. With a few exceptions, every cell in the body of an individual contains a complete copy of their DNA (or genome). For this reason, genetic analysis can be carried out on almost any nucleated cell type (such as lymphocytes or buccal mucosal cells) that can conveniently be collected. The application of DNA analysis now extends well beyond diagnosis of the classic inherited diseases to include, for example, the diagnosis and prognosis of cancer.

This chapter provides a general background to clinical laboratory applications of molecular genetic analysis. The emphasis is on diagnostic techniques with the potential for automation, utilizing polymerase chain reactions (PCRs), since classic techniques such as Southern blot analysis are not widely used in hospital biochemistry laboratories, but tend to be restricted to specialist molecular genetics departments. As far as possible, only a very basic knowledge of molecular biology has been assumed, but several excellent introductions to the topic are available (see Further reading) and a glossary is provided on page 872.

GENES AND GENE EXPRESSION

What is a gene?

A common working definition is that a gene is a sequence of nucleotide bases in DNA that codes for a single polypeptide, but the complexity of genomic organization is such that it is probably unwise to adhere rigidly to any one definition of the gene. Towards the end of the 19th century, it was already accepted that linear groups of ‘invisible self-propagating vital units’ were present in chromosomes. Mendel’s discovery (1865) that inheritance is particulate was rediscovered and publicized at the beginning of the 20th century, and the term gene was introduced to describe Mendel’s ‘particulate elements’ in 1909. By 1911, a specific gene (for colour blindness) had

already been assigned to a particular chromosome (the X chromosome). With the work of Garrod, who first presented his studies on alkaptonuria in 1902, the association of specific diseases with inherited Mendelian traits became established.

Certain stains produce clearly defined bands on chromosomes, so the location of genes is described according to the number of the chromosome on which they are found, whether they are on the long (q) or short (p) arm, and the band number. For example, the location of the α_1 -antitrypsin gene is described as 14q31–32.3, meaning that it is found on the long arm of chromosome 14 in the region of bands 31–32.3. The locations of some of the genes that have been mapped to the X chromosome are shown in Figure 43.1.

After the double helical structure of DNA had been discovered in 1953, the rather abstract concept of a gene became more tangibly associated with a physical structure. Nucleic acids consist of two complementary polymers of nucleotides. Each nucleotide consists of a purine or pyrimidine base, linked to a phosphorylated pentose. In DNA, the pentose is deoxyribose and the bases are adenine (A), guanine (G), cytosine (C) and thymine (T). In ribonucleic acid (RNA), the pentose is ribose and the pyrimidine uracil (U) replaces thymine. Protein coding sequences (exons) are interrupted by non-coding sequences (introns), which are variable in number (up to 50 in collagen genes, for example) and size (up to several thousand base pairs). As a consequence, although the knowledge that three bases code for an amino acid allows us to predict that the coding sequence for an average protein of 400 amino acids will be 1200 nucleotides, the complete gene could be an order of magnitude larger. The boundaries between exons and introns are critically dependent on the GT-AG rule: that is, introns almost invariably begin with GT (or GU in RNA) and end with AG. The structure of a hypothetical gene is shown in Figure 43.2. The promoter region of DNA, which precedes the coding region ('upstream' from the 5' end of the gene, that is, in the opposite direction to transcription) is intimately involved in permitting and regulating expression. Some genes code for RNA (e.g. ribosomal and transfer RNA) that is not translated into protein, and modifications in the process of intron removal can result in one sequence of DNA participating in synthesis of different proteins, so that certain genes can be considered to overlap.

The Human Genome Project

The Human Genome Project (HGP) represents an outstanding piece of multinational cooperation to map the entire human DNA sequence. The project, started in 1990, had several aims, the first of which was to determine the entire base pair sequence. The sequence of 3 billion base pairs was announced in draft form in 2000 and the complete sequence in 2003. It was thought initially that the human genome would consist of approximately 100 000 different coding genes. As the HGP neared completion, it emerged that the actual number would be closer to 30 000.

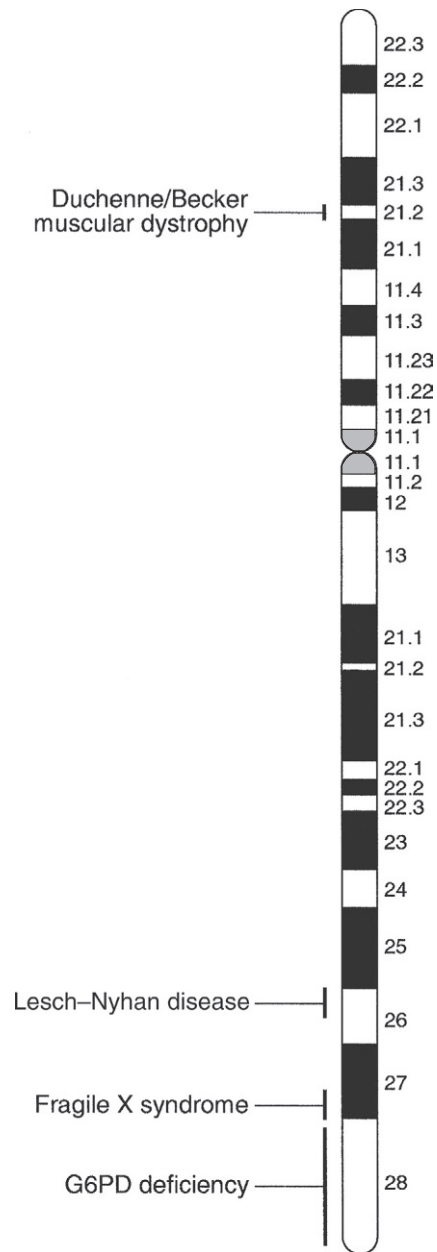


FIGURE 43.1 ■ The mapping of genes to specific sites on the X chromosome. The regions in the X chromosome where the genes associated with Duchenne and Becker muscular dystrophy (the dystrophin gene at p21.2), Lesch-Nyhan disease (the hypoxanthine-guanine phosphoribosyl transferase or HGPRT gene at q26.1-q26.2), fragile X syndrome (the FRAXA gene at q27.3) and glucose 6-phosphate dehydrogenase (the G6PD gene at q28) deficiency are located are shown.

The 'Encode' project

Protein coding sequences and introns account for about 20% of DNA. The function of the remainder is being elucidated but in 2012, the initial findings of the 'Encode' project, which had been examining what had previously been called 'junk' DNA, were published. It appears that the remaining 80% does have a function within the genome, with much of the non-protein coding DNA appearing to code for RNA transcripts that may have other regulatory functions alongside those of gene enhancers and promoters.

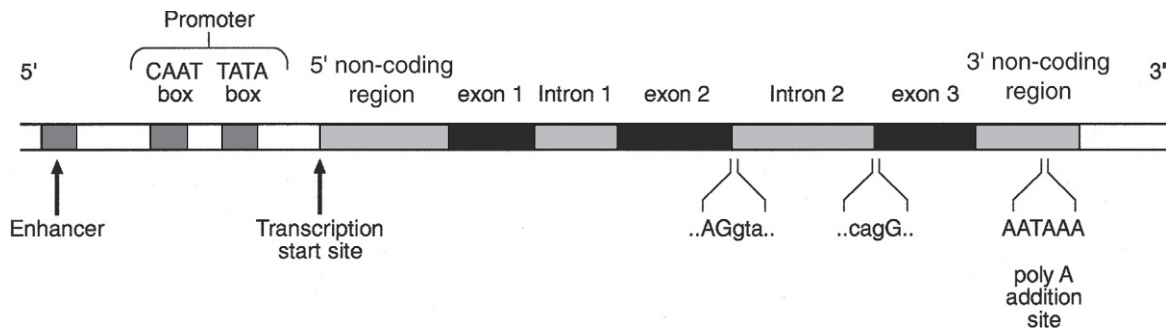


FIGURE 43.2 ■ The structure of a hypothetical gene. Coding sequences (exons) are shown in black, introns and non-coding regions are shaded and regulatory regions are in dark shading. Bases in exon sequences are shown in upper case, while bases in intron sequences are shown in lower case, illustrating the GT-AGT rule for starts and ends of introns.

That the number of genes is much lower than expected appears to be because many genes can perform multiple functions, and it now appears that these functions may be regulated by the remainder of the genome. These discoveries have implications for the investigation and diagnosis of genetic disease, and issues concerned with gene expression will have a growing role in clinical genetics.

Gene expression

The differentiated properties of each cell are determined by the pattern of genes in the cell that are active or dormant. In any one cell, only a small percentage of genes are likely to be actively engaged in directing synthesis of RNA at any one time. Many of these are ‘housekeeping’ genes, which are expressed in virtually all cell types. Much remains to be learned concerning the factors that determine whether or not a gene is expressed, but regulation of gene expression is clearly determined by proteins that interact with DNA.

In higher (eukaryotic) organisms, DNA is found within the nucleus and the mitochondria, although the mitochondria have a very small percentage of the total DNA within the cell and a very small number of genes. Nuclear DNA is complexed with basic proteins (histones) to form chromatin. At regular intervals, DNA is wrapped around complexes of eight histones to form nucleosomes. The copying of DNA into RNA (transcription) is performed by RNA polymerase, which initiates transcription by interacting with the promoter region of a gene. Gene expression is inhibited if nucleosomes cover a promoter region and many factors that regulate transcription probably do so by competing with histones for binding to the promoter region. Within each promoter region, there are several elements that bind specific proteins capable of interacting with RNA polymerase and associated proteins. One of the key proteins in this group binds to the so-called ‘TATA box’ element (actually a TATAAAA or related sequence), which is found in most eukaryotic promoters and is usually located around 30 base pairs (bp) upstream from the transcription start site. Other conserved sequences, such as CAAT, also bind transcription factors and are found within the promoter region. Another class of regulatory sequence in DNA, the enhancer region, binds regulatory molecules, which include the steroid hormone receptors. Enhancer sequences may be some distance from the gene

that they regulate, but the proteins that bind to them may nevertheless interact with the transcriptional apparatus as a result of looping in the DNA molecule.

The initial RNA transcript (pre-mRNA) is modified in several ways before it leaves the nucleus (Fig. 43.3). First, a ‘cap’ structure (7-methylguanosine) is attached to the 5’ end and a sequence of about 200 adenylic residues (poly A) is added at the 3’ end. The non-coding introns are then removed by a two-step splicing mechanism to form a mature messenger RNA (mRNA). Splicing, which takes place within ‘spliceosomes’ (complexes of RNA and proteins), requires cleavage at the 5’ and 3’ ends of the intron and ligation (joining) of the exons. The specific boundary sequences found at splice junctions (see above) act as signals for splicing. Comparison of large numbers of splice junctions reveals these ‘consensus sequences’ to be of the general form AGgta at the 5’ junction and cagG at the 3’ junction (where bases in the intron sequence are in lower case). Given that pre-mRNA molecules can contain up to 65 exons and an intron may consist of thousands of nucleotides, it is remarkable that the correct sites for splicing can be chosen.

Finally, the process of translation involves the activity of ribosomes, transfer RNA and a variety of other molecules which synthesize a protein using the mRNA code as a template. A group of three nucleotides (a codon) specifies an amino acid and most amino acids are coded for by more than one codon (i.e. the genetic code is degenerate). In principle, each RNA sequence can be decoded in three different reading frames, depending on which triplet is chosen as the first codon. In practice, the reading frame is determined by the site of initiation, which always occurs at an AUG codon (AUG codes for methionine, but the initiating methionine is cleaved from proteins in eukaryotic cells). Translation stops at any one of three stop codons (UAA, UAG or UGA). Any subsequent modification of a protein, such as proteolytic cleavage or addition of carbohydrate, is known as post-translational modification.

Mutation, the source of diversity and disease

The accepted terminology, which referred to the ‘normal’ gene in a population as the ‘wild type’, has changed and the terms ‘normal’ and ‘mutant’ (or ‘variant’ if pathogenicity is unclear or questionable) are now preferred.

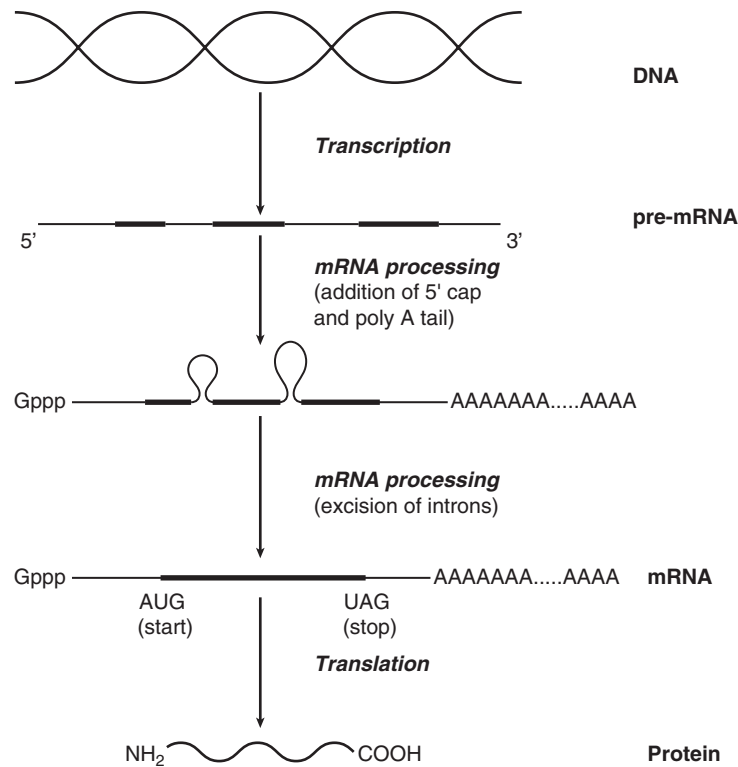


FIGURE 43.3 ■ Transcription and mRNA processing. After transcription, processing of precursor mRNA involves capping, whereby GTP is attached to the 5' end of the mRNA precursors via a 5'-5' triphosphate linkage (i.e. in the reverse orientation to all other nucleotides); addition of around 200 adenylate residues to form a poly(A) tail at the 3' end; and splicing, in which introns are excised and exons spliced together. Translation of mRNA into protein is initiated by ribosomes and transfer RNA at the AUG codon and terminated at one of the stop codons (UAG, UAA or UGA).

However, the genetic constitution of populations is in a constant state of flux with new genes appearing as a result of mutation and deleterious genes being removed by natural selection. Mutations can be broadly divided into those that change the genetic code at a specific location (point mutations or single nucleotide polymorphisms, SNPs) and those that result in gain or loss of genetic material (deletions, duplications and insertions). Point mutations can result from incorrect insertion of a base during DNA replication by DNA polymerase or spontaneous decomposition reactions such as depurination and deamination. Mutagenic chemicals that increase this error rate include those that mimic the natural bases or distort the structure of DNA and those that chemically modify DNA. Ultraviolet light also causes point mutations, particularly by formation of pyrimidine dimers. Point mutations in which a purine is replaced by another purine (e.g. A replaced by G) or a pyrimidine is replaced by another pyrimidine are known as transitions, while replacement of a purine by pyrimidine (e.g. G replaced by C) or vice versa is known as a transversion. Gain or loss of genetic material can result from various errors, including chromosomal breakage and unequal crossing over. Insertion of viral sequences into DNA can also disrupt the genetic code and the rate of spontaneous chromosomal breakage can be markedly increased by ionizing radiation. Whenever the number of bases deleted or inserted is not a multiple of three, the reading frame of the mRNA is altered (frameshift mutation) and the RNA sequence subsequent to the mutation becomes nonsense.

Accumulated damage to DNA would rapidly overwhelm the organism, but repair mechanisms recognize and repair damaged DNA so that fewer than 1 in every 1000 accidental base changes results in a stable mutation. It is estimated that stable point mutations are acquired at a rate of about 1 in every 10^9 base pairs during each cell generation. Consequently, an average gene of about 10^3 coding base pairs is likely to acquire a mutation once in every 10^6 cell generations. As might be expected, individuals with inherited defects in the enzymes responsible for DNA repair are markedly more susceptible to the effects of environmental mutagens.

A significant proportion of germline point mutations are thought to be caused by modification of methylated cytosine residues. DNA methylation, restricted in eukaryotic cells to cytosine residues, usually at CpG dinucleotides (CpG denotes C-phosphate-G in a linear sequence, in distinction from a CG base pair) is not present in all such organisms but is thought to play an important role in ensuring stable inheritance of expression patterns when cells divide. Spontaneous deamination of the 5-methylcytosine creates thymidine, and problems can then arise when the normal guanine on the complementary strand becomes an adenine and the mutation cannot be detected by DNA repair mechanisms. Consequently, methylation of cytosine can produce mutation 'hot spots' (i.e. sequences associated with an unusually high frequency of mutations or recombination).

Some mutations may have no effect on the structure of a protein – either because the genetic code is degenerate

and the new sequence codes for the same amino acid, or because some amino acids in a protein can be substituted without producing any significant effect on the function of the protein. However, some apparently 'silent' mutations may have an effect on the protein product, not because of the actual base change involved, but rather through an effect on splicing by activating cryptic splice sites or destroying splice enhancers. Mutations that change the three-dimensional structure of a protein and so alter its function or stability may do so by a variety of mechanisms (Fig. 43.4).

A Missense

 Ile	Asp	Glu	Lys	Gly
Normal: ATC	GAC	GAG	AAA	GGG
Mutant: ATC	GAC	AAG	AAA	GGG
 Ile	Asp	Lys	Lys	Gly

B Nonsense

 Arg	Leu	Gln	Glu	Glu
Normal: CGA	CTG	CAG	GAG	GAG
Mutant: CGA	CTG	TAG	GAG	GAG
 Arg	Leu	stop			

C Splice site mutation

			← exon 12	intron →	
Normal: ATT	AAC	A	gtaagt
Mutant: ATT	AAC	A	ataagt

D Deletion

 Ile	Ile	Phe	Gly	Val
Normal: ATC	ATC	TTT	GGT	GTT
Mutant: ATC	ATT	GGT	GTT	
 Ile	Ile	Gly	Val	

E Insertion

		Lys	Gly	Gly	Tyr	Lys
Normal: AAG	GGG	GGC	TAT	AAA
Mutant: AAG	GGG	GGG	CTA	TAA
		Lys	Gly	Gly	Leu	Stop

F Expansion of triplet repeat

Normal: GCG	(CGG) ₅₋₅₄	CTG
Mutant: GCG	(CGG) _{>200}	CTG

FIGURE 43.4 ■ Examples of mutations. (A) G to A transition in the α_1 -antitrypsin gene, a missense mutation resulting in substitution of glutamic acid by lysine at position 342, producing the Z variant associated with α_1 -antitrypsin deficiency. (B) Transversion of C to T in the steroid 21-hydroxylase gene, converting the codon for glutamine to a stop codon, one of the mutations causing congenital adrenal hyperplasia. (C) G to A transition in the 5' splice site of intron 12 in the phenylalanine hydroxylase gene, resulting in deletion of exon 12, the most frequent cause of phenylketonuria in Caucasians. Bases in intron sequences are shown in lower case. (D) Deletion of three bases in the cystic fibrosis transmembrane conductance regulator (CFTR) gene resulting in deletion of phenylalanine at position 508, the most frequent mutation causing cystic fibrosis in Caucasians. (E) Insertion of a G in the hypoxanthine-guanine phosphoribosyl transferase gene, a frameshift mutation that results in the Lesch-Nyhan syndrome. (F) Amplification of a CGG triplet repeat in the FMR-1 gene, which causes the fragile X syndrome.

By convention, the DNA strand that has the same sequence as mRNA (except that it possesses T instead of U) is represented. This strand is known as the coding strand, although it is the 'anticoding' strand that is complementary to mRNA and therefore provides the template for mRNA synthesis.

Some amino acid changes (missense mutations), such as that which produces the Z variant of α_1 -antitrypsin (Fig. 43.4A), can have a profound effect on the processing or function of a protein. Some mutations (nonsense mutations) create or destroy codons for the start or stop signals of translation so that a protein of abnormal length is produced (Fig. 43.4B). Mutations at splice sites (Fig. 43.4C) frequently result in production of abnormal mRNA that is unstable. Deletion of three bases removes the codon for a single amino acid without altering the reading frame, as occurs in the most common mutation causing cystic fibrosis (Fig. 43.4D). Insertion (Fig. 43.4E) or deletion of any number of bases that is not a multiple of three will alter the reading frame so that the message becomes garbled. Amplification of triplet repeat sequences (Fig. 43.4F) has been identified as the basis of several inherited diseases. Occasionally, mutations affect regulatory regions of DNA so that the amount of protein produced is altered. Although mutations are most frequently either neutral or deleterious, rare mutations will alter the function of a protein in such a way that the fitness of an individual is improved, so contributing to evolution. Inheritance of the mutations that have accumulated in our ancestors, whether they be advantageous, neutral or deleterious, is what constitutes our individuality.

An individual inherits two copies of each chromosome (one maternal and one paternal). On each chromosome, the sequences at each site, or locus, are known as alleles. If the two alleles are identical, the individual is said to be homozygous at that locus, while if the alleles are different, the individual is said to be heterozygous for each allele. As will be seen later, genetic disease is usually heterogeneous, so that an individual said to be homozygous for a deleterious gene may be found, when studied at the molecular level, to carry a different mutation in each allele (i.e. is a compound heterozygote). When the prevalence of a mutant allele becomes more common in a population than could be maintained by new mutations alone (generally taken to be when more than 1% of the population carry the allele), there is said to be polymorphism. Many proteins in blood (e.g. haptoglobin) and on cell surfaces (e.g. human leukocyte antigen, HLA) are polymorphic.

The classic inherited diseases result from single gene defects and more than 6000 inherited diseases likely to be associated with defects in single genes have already been identified. Inherited diseases could, in theory, result from mutations in any one of the human genes – the only limitation being that the structure of some gene products is so critical that any mutation will not be compatible with life. Most of the more common diseases that afflict western society, including most cases of diabetes, atherosclerosis and hypertension, are the result of interaction between the environment and polygenic factors (i.e. they are determined by interactions between

several genes). Molecular analysis of the polygenic diseases is considerably more difficult than analysis of single gene defects, but alleles that predispose individuals to development of these diseases are being identified. Both single gene defects and most multifactorial/polygenic diseases arise from mutations in the nuclear DNA but genetic diseases also arise from mutations in mitochondrial DNA and from chromosomal abnormalities. Mitochondrial DNA is extranuclear and shows almost complete maternal inheritance. Chromosomally inherited disorders include the trisomies, where faulty meiosis allows two copies of a chromosome to be present in a gamete, leading to three copies in the embryo. Trisomy of chromosome 21, for example, is responsible for Down syndrome.

If new mutations occur in germ cells, then they may give rise to an inherited disease in the next generation. The effects of mutations on non-germ cells, or somatic cells, will depend both on the gene affected and on the state of differentiation of the cell affected. The ageing process is likely to be one result of accumulated mutations in somatic cells and the central role of mutations for the development of cancer has become clearer during the last few years.

Genesis of an individual: the formation of gametes

An individual's genotype is determined at the time of fertilization, when the chromosomes of the gametes (i.e. the male sperm and female egg) are combined. The formation of gametes (gametogenesis) is particularly relevant to an understanding of the detection of inherited disease because it is at this stage that 'shuffling' of genes occurs.

Normal cell division, or mitosis, involves a simple copying of each chromosome, with one identical copy being passed on to each daughter cell. To avoid a doubling in the number of chromosomes in each generation, gametogenesis involves a reduction (by half) of the chromosome number during two specialized cell divisions known as meiosis. Since the chromosome complement of parent and offspring must be equivalent, the reduction in chromosome number cannot be arbitrary: parental contributions must be equal and equivalent. This requirement can be met because each somatic cell of an individual is diploid, containing corresponding (homologous) pairs of chromosomes, one derived from the mother and the other from the father. Meiosis consists of two cell divisions: in the first, after duplication of DNA, pairing of homologous chromosomes occurs, and to ensure that each gamete receives just one member of each homologous pair, the duplicated paternal chromosome is distributed to one, the duplicated maternal chromosome to the other. The assortment between the two cells appears to be random so that each cell acquires some maternal and some paternal chromosomes. The second cell division is like ordinary mitosis, except that it is not preceded by duplication of chromosomes. As a consequence, the gametes produced are haploid, with half the normal number of chromosomes.

At first sight it might be expected that a chromosome would be transmitted from one generation to the next as an intact unit and that two genes on the same chromosome would always be inherited together. The fact that this is not so is a result of events that occur during the first meiotic division, which have important consequences for genetic analysis. As chromosomes pair prior to the first division, crossovers (or chiasmata) occur by breakage and rejoining between the chromatids of homologous chromosomes resulting in recombination (Fig. 43.5). The recombination fraction is a measure of the genetic, rather than the physical, distance between two genes (or loci). The recombination fraction for two loci can never be more than 0.5 as the resulting chromatids can only ever be recombinant or non-recombinant, no matter how many crossovers have occurred between the loci. In simple terms, linkage of two genes (i.e. a tendency to be inherited together) occurs only when the genetic distance separating them is sufficiently short to make crossover between them unlikely. The association of two genes on separate chromosomes is random but association of genes on the same chromosome is not, as it is known that crossovers do not occur at random. A process known as interference prevents a chiasma forming if one already exists nearby. When genes are associated more frequently than would be predicted by chance, they are said to be in linkage disequilibrium.

In females, gametogenesis is initiated during fetal development and, at birth, the germ cells are in a phase of arrested maturation of the first stage of meiosis, which is not completed until ovulation. The increased risk of chromosomal abnormalities in older mothers may be

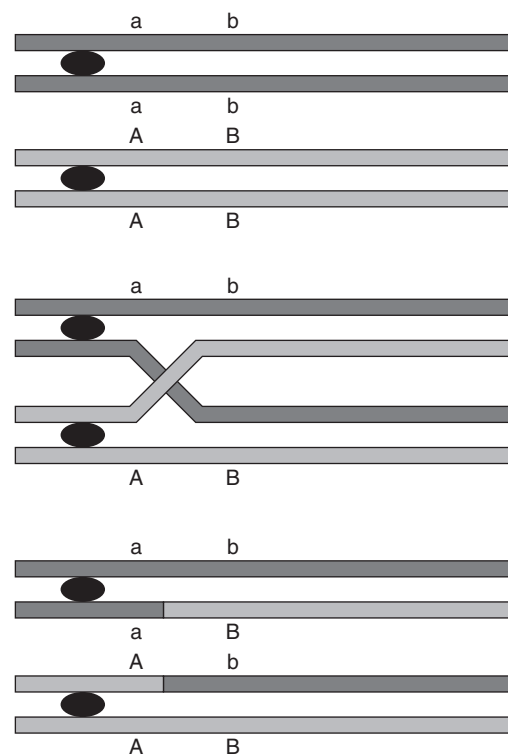


FIGURE 43.5 ■ Crossover and linkage. Exchange of alleles as a result of crossover between homologous chromosomes. Linkage between alleles a and b (or A and B) occurs if they are sufficiently close that crossover is unlikely to occur between them.

explained by the fact that completion of meiosis occurs only after ovulation, when the second stage of meiosis (which is similar to ordinary mitotic division) occurs and during which fertilization can take place. This may happen up to 50 years after formation of the germ cells. In males, sperm production continues from the time of sexual maturity into old age and the large number of cell divisions involved is probably the cause for the increased number of new single gene mutations that seem to occur in the children of older men.

Genes in families and populations

Mendel introduced the concepts of dominant 'characters' (or traits), which are transmitted without change, and recessive traits, which become latent after cross-fertilization. In modern terms, an allele coding for a dominant trait can be said to manifest its phenotype in the heterozygous state (i.e. only one copy is required for its effects to be apparent), while an allele for a recessive trait expresses its phenotype only in homozygotes. An individual who is heterozygous for an autosomal recessive condition is described as a carrier. The types of family pedigree associated with autosomal dominant and recessive genes are illustrated in Figure 43.6A and B. Autosomal dominant conditions affect males and females equally, and affected individuals who are heterozygous for the abnormal allele will transmit it to half of their offspring. Autosomal recessive disorders occur

in individuals whose parents are both carriers for the mutant gene. The risk for such patients of having affected children is 25% and the probability of any child they have being a carrier is 50%. Dominant diseases are often associated with genes coding for proteins that have structural, carrier or receptor functions, while genes coding for enzymes are often associated with recessive disorders. The explanation for this is probably that the activity of most enzymes is considerably greater than necessary for normal metabolism, so loss of up to half of the normal activity is of little consequence. Inheritance is somewhat different for alleles on X chromosomes. Dominant X-linked diseases will affect both males and females, but an X-linked recessive disease will be manifest only in males, who have only one X chromosome (Fig. 43.6C). In females, during the very early stages of embryogenesis, one or other X chromosome is inactivated in every cell; consequently, females can be carriers but will usually only suffer from an X-linked recessive disease if they are homozygous. However, sometimes, in some disorders, the inactivation is not random but occurs in such a way that only the normal chromosome is inactivated, with the result that only mutant alleles are expressed in critical tissues ('skewed' X-inactivation) and the individual becomes a manifesting heterozygote. This has been reported in several X-linked recessive disorders including, for example, Duchenne muscular dystrophy.

When both alleles are expressed in a heterozygote, each producing its phenotype independently, inheritance is said to be codominant. This type of inheritance is seen most clearly when phenotypes are determined by immunological or biochemical tests – for example, in testing for blood groups and in restriction fragment length polymorphism (RFLP) analysis and, more recently, in the growing field of pharmacogenetics (see later), where it is possible to distinguish both alleles at each locus.

Of the inherited diseases currently recognized, the overwhelming majority are due to problems in nuclear DNA and are autosomal dominant, autosomal recessive or sex linked. However, it is now known that some diseases are associated with the small amount of DNA that is present in mitochondria. These diseases are maternally inherited, since the mitochondria in the fertilized egg are of maternal origin.

The relative frequency of inherited disease varies markedly between populations: for example, cystic fibrosis and α_1 -antitrypsin deficiency are associated primarily with northern Europeans, while red cell disorders (thalassaemia, sickle cell anaemia and glucose 6-phosphate dehydrogenase deficiency) are found primarily in people of Mediterranean, Oriental or African origin, and Tay-Sachs disease is found primarily in Ashkenazi Jews. In some populations, the prevalence of an inherited disease is due to a 'founder effect', as with variegate porphyria in South Africans, which can be traced to a single couple who emigrated from Holland in the 1680s. Autosomal recessive diseases that are particularly widespread in larger populations are likely to represent a balanced polymorphism in which disadvantage to homozygotes is balanced by an advantage to the larger number of heterozygotes. With some red cell disorders (e.g. HbS, the

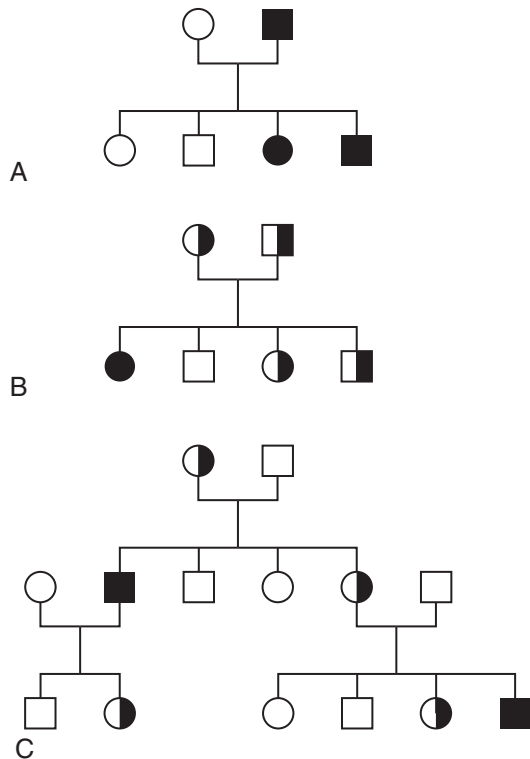


FIGURE 43.6 ■ Patterns of inheritance in families. Inheritance of (A) a dominant condition, (B) a recessive condition and (C) an X-linked recessive condition. Squares represent males and circles represent females. Open symbols represent normal individuals, while fully shaded symbols refer to affected individuals and half-shaded symbols represent carriers.

sickle cell trait), the balanced polymorphism is clearly a response to the environment, in this case with heterozygosity conferring resistance to malaria caused by *Plasmodium falciparum*.

In a large population, the relative frequencies of different alleles tend to remain constant and a simple mathematical formula allows calculation of the frequency of different genotypes. If two alleles, A and a, occur at a given locus and their frequencies are p and q , respectively, then:

$$p + q = 1$$

(since one or other allele must be present,
their sum must be 1)

It can be shown that the genotypes AA, Aa and aa have frequencies p^2 , $2pq$ and q^2 , respectively (the Hardy-Weinberg law). Use of this law allows simple calculation of carrier frequencies for autosomal traits. For example, if the homozygote frequency (q^2) for cystic fibrosis is about 1 in 2500, q is $1/50$, p is $1 - q$ or $49/50$ (~ 1) and the heterozygote frequency ($2pq$) is about $1/25$.

The variable expression of genetic disease

Several factors dictate that each genetic disease is associated with symptoms of quite variable nature and severity.

It is often naively assumed that all mutations of a particular gene will have identical consequences for the organism, but this is far from being true. Different mutations in a given gene are quite likely to give rise to different phenotypes, just as a particular phenotype can result from many different mutations in the same gene or even in different genes. Since many different mutations that have deleterious effects on a particular gene are present in most populations, it is not surprising that molecular analysis often reveals that individuals who have been described as homozygous are in fact compound heterozygotes, that is, they are affected by two different deleterious alleles.

Because the haemoglobin gene has been investigated in detail, it provides a good illustration of the complexity of genetic disease. Several hundred abnormal haemoglobins have been identified. The majority have amino acid substitutions resulting from single base changes and the consequences range from complete absence of protein to variant haemoglobins with function indistinguishable from that of normal (see Chapter 29). As study of other genes progresses, similar complex arrays of mutations of every imaginable type and with differing consequences are being discovered. With the introduction of screening for phenylketonuria (PKU), it soon became apparent that the disease is heterogeneous. It has been shown that, in PKU, phenotypic heterogeneity is related to the level of phenylalanine hydroxylase activity expressed in each patient which, in turn, is determined, in classic PKU, by the particular mutations that are present in the two alleles of the phenylalanine hydroxylase (PAH) gene. However, not all cases of neonatal hyperphenylalaninaemia are due to PAH deficiency. Benign hyperphenylalaninaemia arises from a transient liver immaturity and does not lead to the disease. Two rare causes of PKU are

deficiencies in the enzymes dihydropteridine reductase and dihydrobiopterin synthetase, leading to PKU with a severe phenotype. Glucose 6-phosphate dehydrogenase (G6PD) deficiency, which is an X-linked recessive condition estimated to affect as many as 500 million people worldwide, has also proved to be extremely heterogeneous at both the phenotypic and molecular levels. Some variants appear to have no clinical consequences, while at the other extreme, severe defects in G6PD cause hereditary non-spherocytic haemolytic anaemia (see Chapter 27). Intermediate defects are associated with haemolytic anaemia only in the presence of precipitating factors (e.g. infection, ingestion of fava beans or certain drugs). Analysis of the protein had indicated the existence of about 400 variants of G6PD and a similar number of mutations have now been described, but many of these appear not to cause disease. Many of the mutations causing the most severe disease are clustered near the carboxy end of the enzyme in the region of the putative NADP binding site.

A further level of heterogeneity in genetic disease occurs as a result of varying degrees of penetrance and expressivity. Penetrance refers to the degree to which the mutation causes disease. Thus, in some disorders, the presence of the mutant gene is disease-causing in some individuals but not others, demonstrating variable penetrance, whereas in diseases that are fully penetrant, the presence of the mutation will always lead to the disease phenotype. For example, the C282Y (Cys282Tyr) mutation in the *HFE* gene (causing substitution of the cysteine residue at position 282 of the human haemochromatosis protein by tyrosine), causes haemochromatosis in some individuals but results in a completely normal phenotype in others. Expressivity is a slightly different aspect of a gene's effect and refers to the presence of variable phenotypes arising from the same mutation. It can be age-related or determined by the environment (as when drugs such as barbiturates precipitate attacks of acute intermittent porphyria). An example of a disorder demonstrating variable expressivity is Waardenburg syndrome, where the 'full' syndrome includes several phenotypic features (such as different coloured eyes, a white forelock or deafness) but where, within a single affected family, different individuals may have only one feature, which is not always the same one within that family. Variable expressivity and penetrance tend to be features of dominant, rather than recessive, conditions.

Another aspect of inheritance, known as imprinting, may help to explain processes such as variable penetrance and expressivity. Contrary to the assumptions of classic genetics, it now appears that expression of some genes depends on whether they are of maternal or paternal origin. The molecular mechanism of imprinting, which is likely to occur during meiosis, involves DNA methylation, which 'marks' certain genes and ensures that they are preferentially expressed in the next generation. If such a gene is imprinted in the maternal line, it will continue through a woman's daughters but not through her sons, although both may be affected. The converse is true for paternally imprinted genes, which will be transmitted through

sons but not through daughters. Consequently, if an imprinted gene contains a deletion, offspring will not show expression of the gene product even in the presence of a normal gene on the opposite chromosome, as this will be 'switched off'.

Prader–Willi and Angelman syndromes are good examples of disorders arising due to abnormalities in a region carrying imprinted genes (chromosome 15 at 15q12.). Prader–Willi syndrome, features of which include hypotonia and hyperphagia, is produced by deletion of the paternal alleles at 15q12. Angelman syndrome, which is associated with ataxic movements and seizures, is also associated with deletion of 15q12 but, in this case, of the maternal allele. In some cases of Prader–Willi syndrome, rather than deletion of the paternal alleles, loss of the paternal chromosome occurs together with maternal isodisomy (two copies of the same allele from the mother) or heterodisomy (one copy of each maternal allele). There are several possible mechanisms by which two alleles can be inherited from one parent. Trisomies, for instance, usually result in spontaneous abortion, but if one chromosome is then lost there is a one in three chance that resulting cells will have a normal complement of chromosomes but with one pair of chromosomes derived from a single parent. The interesting possibility that a recessive disease can be inherited from one carrier parent then arises. This unusual mode of inheritance has been demonstrated in some patients with cystic fibrosis, but it is not yet clear how frequently it occurs in this or other diseases.

THE TECHNIQUES OF GENETIC ANALYSIS

Detection of specific sequences in DNA

Analysis of DNA is heavily dependent on the availability of techniques to identify specific nucleotide sequences. Fortunately, the function of DNA has resulted in the evolution of proteins capable of recognizing specific DNA sequences and it is inherent in the structure of DNA that one strand should recognize and bind (hybridize) specifically to its complementary strand. Most of the techniques currently used in DNA technology exploit one or other of these properties.

Use of proteins that recognize DNA sequences: restriction endonucleases

Without restriction enzymes, much of the molecular biological analysis carried out in the last 30 years would not have been possible. These enzymes are widespread in bacteria: over 3000 having been recognized so far, of which around 600 are available for commercial/analytical use. Each enzyme is named after the species of bacteria in which it was found (e.g. *EcoRI* from *Escherichia coli*), and in which it probably fulfils a defensive function, cleaving molecules of foreign DNA. The usefulness of these enzymes derives from the fact that they do not cleave DNA at random, but recognize and cut specific nucleotide sequences. The most commonly used restriction enzymes recognize sequences of 4–6 nucleotides that have a two-fold axis of symmetry

and are therefore said to be palindromes (i.e. the sequence reads the same on the complementary strand) (Fig. 43.7A). Digestion of DNA by a particular enzyme provides reproducible fragments whose size will depend on the frequency with which the enzyme recognition site occurs. On average, a 4-bp site occurs every 256bp and a 6-bp site every 4096bp. While some enzymes (e.g. *HaeIII*) cut in such a way that 'blunt ends' are produced, others (e.g. *EcoRI*) cut asymmetrically so that 'sticky ends' are left, which are extremely useful for reannealing fragments to produce recombinant DNA (Fig. 43.7B).

In addition to allowing reproducible cleavage of DNA to a manageable size, restriction enzymes are also valuable tools for analysing molecular diversity and identifying the individuality of DNA sequences. Differences in DNA sequences between individuals may create or destroy sites for restriction enzymes (i.e. there is polymorphism of restriction sites). Thus, the distance between restriction sites will often differ between individuals and between the maternal and paternal strands of DNA. The pattern of restriction sites can therefore provide a 'signature' for each individual strand of DNA. The different populations of DNA fragments produced on digestion by an enzyme are known as restriction fragment length polymorphisms or RFLPs.

Restriction enzymes have more recently acquired a new use in preparing genomic DNA for the various techniques employed in 'next generation' sequencing.

Hybridization: probes and the polymerase chain reaction (PCR)

A probe is a sequence of DNA (or RNA) that has been labelled in order to identify complementary base sequences by molecular hybridization. The two strands of DNA can be dissociated ('denatured' or 'melted') in various ways, such as by heating or addition of alkali. Denaturation for a given fragment of DNA occurs at a specific temperature, and the temperature at which 50% of the duplex is dissociated is known as the T_m . When the temperature is lowered to just below the T_m , hydrogen bonds begin to reform between complementary bases, a process known as annealing or renaturation. If the probe and target DNA are mixed before reannealing is allowed to occur, the probe can be used to 'find' its complementary sequence. The conditions under which reannealing occurs (in particular salt concentration and temperature) determine the degree of stringency of the hybridization.

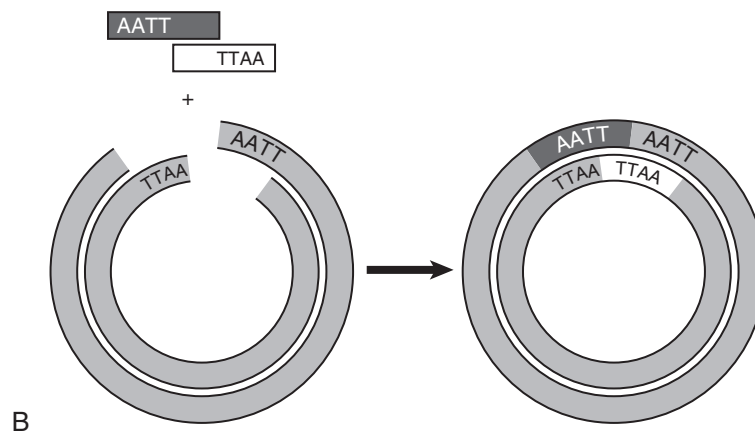
Probes can be used to detect their complementary sequences, traditionally after DNA (or RNA) fragments have been separated by electrophoresis. Digestion of genomic DNA with a restriction enzyme will produce a million or so fragments of different sizes and electrophoresis has the great advantage of allowing simple determination of the size of fragments detected by probes. The now classic technique of Southern blotting involves the transfer of electrophoretically separated bands of DNA to a sheet of nitrocellulose or nylon. Complementary DNA sequences are then detected by hybridization with labelled probes. The technique was originated by Dr (later Professor) E M Southern. Subsequently, the terms Northern and

Some restriction enzymes and their target sequences

Microorganism	Enzyme	Sequence
<i>Bacillus amyloliquefaciens</i>	Bam HI	$\begin{array}{c} \dots G \uparrow G A T C C \dots \\ \dots C C T A G \downarrow G \dots \end{array}$
<i>Escherichia coli</i>	Eco RI	$\begin{array}{c} \dots G \uparrow A A T T C \dots \\ \dots C T T A A \downarrow G \dots \end{array}$
<i>Escherichia coli</i>	Eco RV	$\begin{array}{c} \dots G A T \uparrow A T C \dots \\ \dots C T A \downarrow T A G \dots \end{array}$
<i>Haemophilus aegyptius</i>	Hae III	$\begin{array}{c} \dots G G \uparrow C C \dots \\ \dots C C \downarrow G G \dots \end{array}$
<i>Haemophilus influenzae</i> Rd	Hind III	$\begin{array}{c} \dots A \uparrow A G C T T \dots \\ \dots T T C G A \downarrow A \dots \end{array}$
<i>Providencia stuarti</i>	Pst I	$\begin{array}{c} \dots G T G C A \uparrow G \dots \\ \dots G \downarrow A C G T C \dots \end{array}$

A

Use of cohesive ends for production of recombinant DNA



B

FIGURE 43.7 ■ Restriction enzymes and recombination. (A) Arrows indicate how specific DNA sequences are cleaved by restriction enzymes. (B) Cohesive ends produced by restriction enzymes can be used for annealing of DNA sequences, which can then be joined by a ligase to form a recombinant DNA molecule.

Western blotting have been used for processes in which RNA or proteins, respectively, are transferred.

In the past, all DNA analysis required some technique of visualizing the products of the reaction, and with probe hybridization this involves labelling the probe in some way. ³²Phosphorous radioactive labelling, although still in use for blotting protocols, was superseded by the use of fluorescent dyes, particularly for gene sequencing, a system which allowed the development of large, high-throughput, DNA analysers, as well as techniques such as real-time PCR. However, some of the technologies employed in 'next generation' sequencing systems require neither gels nor dyes, as will be discussed later. Some of the newer systems allow for an electrical signal to be generated when hybridization of probe to target

occurs, which may eventually lead to the development of point-of-care systems.

Probes, generally of a few thousand bases, may be sequences cut from genomic DNA or they may have been produced by making complementary DNA (cDNA) to an mRNA species. The latter procedure uses the enzyme reverse transcriptase, which transmits genetic information in the 'reverse' direction, that is, from RNA to DNA. Genomic and cDNA sequences differ, particularly in the absence of intron sequences from the latter. The tolerance of probes for mismatching of base sequences will depend on their size and on the stringency of hybridization. Shorter oligonucleotide probes are often more useful for direct identification of point mutations since

conditions can be chosen such that hybridization occurs only when there is complete complementarity between probe and target. Because hybridization of a relatively smaller proportion of bases in the probe is required, larger genomic or cDNA probes will usually recognize corresponding sequences of DNA from different individuals in a population or even different but related genes.

Cloning, where the sequence of interest was 'grown' in bacteria such as *E. coli* after being inserted into the bacterial genome using bacteriophage viruses, has been largely superseded as a method of producing probes by the ability to create synthetic oligonucleotide sequences, although cloning as a technique is still employed in areas of research.

Without doubt, the commonest use for synthetic oligonucleotides has been as 'primers' in the polymerase chain reaction (PCR). After its introduction in 1985, PCR supplanted many of the more tedious techniques of molecular biology and opened up completely new possibilities. Essentially, a means of cloning DNA without the need for vectors or bacteria, PCR uses the enzyme DNA polymerase to copy DNA. To do this, the enzyme needs two oligonucleotide primers that are complementary to sequences flanking the region of interest in the target DNA, with one on each strand (Fig. 43.8). Computer and web-based programs are available for designing primers (usually 20 or more bases) in order to choose sequences likely to be most suitable for the PCR reaction and to maximize specificity by ensuring that the

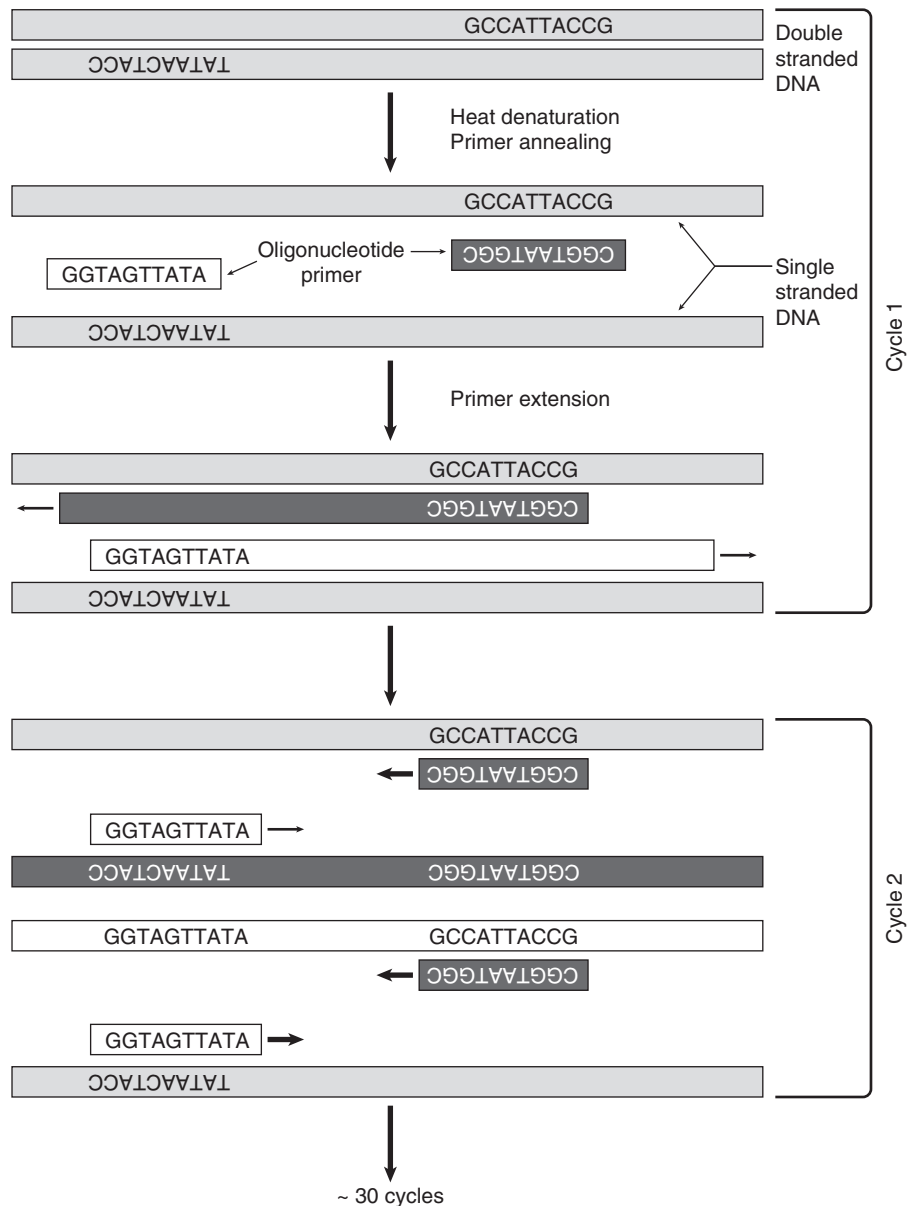


FIGURE 43.8 ■ The polymerase chain reaction. After denaturation of DNA, primers anneal to complementary sequences. During the first cycle, a heat-stable DNA polymerase (usually *Taq*) initiates synthesis of two new strands. After a further denaturation cycle, primers anneal to the newly synthesized DNA as well as the original sequences and four new strands are synthesized (cycle 2). With an exponential increase in the number of DNA strands, after around 30 cycles the sequence will have been amplified about one million times.

complementary sequence is virtually unique in the genome. The target sequence is amplified exponentially by repeated cycles of enzymatic copying. In the first cycle, double-stranded DNA is denatured by heating to between 92°C and 96°C for 5 min, and then cooled to (usually) between 55°C and 60°C, so that the oligonucleotide primers can anneal to their complementary sequences in the target DNA. For optimal specificity, the highest annealing temperature possible is used to minimize extension of primers bound non-specifically. Extension of the primers by DNA polymerase, using added nucleotides and the DNA target as a template, is then allowed to take place. Use of a heat-stable DNA polymerase, which can withstand the heating cycles, avoids the necessity of adding fresh enzyme at each cycle and allows extension to take place at a high temperature (72°C). After heat denaturation, excess primers can then anneal to the newly synthesized DNA as well as the original DNA strands and the process is repeated. In 30 cycles, amplification of over one million-fold can be achieved. Originally, the only enzyme available for PCR was that from *Thermus aquaticus* (*Taq* polymerase), but more have now been identified and developed, making amplification of much longer sections of genomic DNA possible.

Stringent control of conditions, such as the concentrations of magnesium and nucleotides in the reaction mixture, have usually been needed to maintain the specificity of a PCR although conditions that maximize polymerase fidelity may reduce PCR efficiency. Where such manipulation was necessary, it was predominantly dependent on the sequences involved, especially that of the primers. However, the ability to amplify longer sections of DNA means that primers can be 'picked' to suit the conditions required, allowing the use of universal 'master mixes' and thus a greater standardization of PCR assays. PCR is so sensitive that it has been used to amplify the DNA from a single cell and it can be used with samples obtained from materials as diverse as ancient mummies, fossils, hair follicles, preimplantation embryos and fixed pathological specimens. The large quantities of amplified DNA produced by PCR, which can be detected using a variety of visual labels, have eliminated the need for radiolabelling. The extreme sensitivity of the technique is also the source of one of its disadvantages – contamination by extraneous DNA (e.g. from the operator) can create havoc, so that strict precautions must be taken to avoid artefacts.

Non-isotopic fluorescent labels are employed in the widely used technique of real-time PCR. There are several versions of this technique but the basic principle is that a PCR reaction is followed, in real time, by monitoring the signals produced by dye-labelled probes that bind to the accumulating products. Various modifications to the technique allow real-time PCR to be used for quantitation of a target sequence (hence, its other, abbreviated name of qPCR) and determining dosage (the number of copies of a gene in a cell) and gene duplication as well as for mutation detection. Quantitative analysis has become increasingly important as a tool for microbiologists for determining such things as viral load in samples from patients suffering from diseases such as human immunodeficiency virus infection.

Not to be confused with real-time PCR, reverse transcriptase PCR (rtPCR) uses purified RNA as the start point and uses the enzyme reverse transcriptase to produce cDNA, which can then be amplified using conventional PCR and used to examine the protein expressed by the gene of interest. Potentially, this allows investigation of both the effect of any sequence variants in a gene on its protein product, and the possible phenotypic consequences.

Detection of mutations

The search for mutations causing disease involves either the detection of a previously identified, known mutation or the screening of the gene to search for an unknown mutation. In clinical samples, the process will usually begin with the former and proceed to the latter if none of the former is found. There are now a variety of methods available for mutation detection and these can be divided into manual, low-throughput techniques and automated, high-throughput ones. The basic principles of the manual methods often form the basis of the automated ones and it is for this reason that several of the manual techniques are described here as they demonstrate principles of analysis and the properties of DNA that are exploited, however it should be acknowledged that in most large genetics laboratories these have largely been superseded by automated high-throughput systems.

Detecting known mutations

The principles of several manual techniques for detecting known mutations are shown in Fig. 43.9. The allele-specific oligonucleotide (ASO) technique utilizes synthetic oligonucleotide probes (about 19 residues) – one that is complementary to the normal and one to the mutant allele, with the site of mutation in the middle of the complementary region (Fig. 43.9A). Under appropriate conditions of hybridization, the oligonucleotides will bind only when there is complete complementarity. This technique was first used for detection of the β^s allele of the β globin gene that is responsible for sickle cell anaemia. Although ASO testing can be performed directly on genomic DNA, it is more easily carried out after prior amplification of the region of interest using PCR. If a mutation (or the corresponding normal sequence) happens to be a target site for one of the many restriction enzymes, then testing for the presence or absence of a restriction site by restriction endonuclease allele recognition can be used to analyse genotypes (Fig. 43.9B). This technique was also first used for detection of the β^s allele. Again, initial PCR amplification simplifies the procedure.

Another method for identifying point mutations exploits the specificity of the primers used for PCR: the amplification refractory mutation system (ARMS) (Fig. 43.9C), also known as allele-specific amplification, is based on the observation that oligonucleotides that are not exactly complementary to the target DNA sequence frequently do not function as primers in PCR. The method uses two forward primers, one which is complementary to the normal sequence and one which

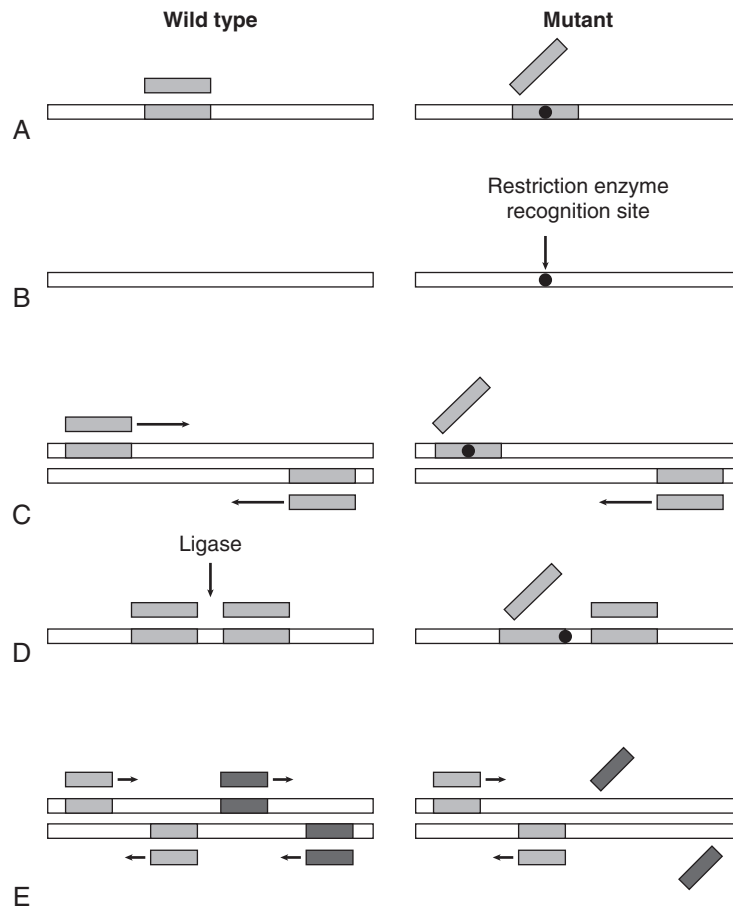


FIGURE 43.9 ■ Some techniques for detecting known mutations. (A) Allele-specific oligonucleotides. A synthetic oligonucleotide complementary to the normal sequence will bind to wild-type, but not mutant, DNA (while an oligonucleotide complementary to the mutated sequence will bind to mutant DNA but not wild-type DNA). (B) Restriction endonuclease allele recognition. A mutation may create a recognition site for a restriction enzyme which is absent in the wild-type sequence. Consequently, digestion with the enzyme would cleave mutant DNA into two fragments, but would have no effect on wild-type DNA (similarly, a mutation may destroy a recognition site for a restriction enzyme so that wild-type DNA is cleaved, but not mutant). (C) Amplification refractory mutation system (ARMS). Primers, one of which spans the site of mutation, are designed for a PCR reaction. Primers complementary to wild-type sequence amplify normal DNA, but not mutant DNA (for confirmation, a second set of primers which are complementary to the mutated sequence is also designed so that a positive reaction is obtained with mutant DNA). (D) Ligase-mediated allele detection. Oligonucleotides flanking the site of mutation are joined by a ligase only if they are complementary to the DNA sequence. One of the oligonucleotides corresponds to a constant region, while the other is complementary to either wild-type or mutant DNA. (E) Multiplex PCR for detection of deletions. Simultaneous PCR reactions for amplification of two or more sequences detect any deletions which remove recognition sites for PCR primers.

is complementary to the mutant sequence, in combination with a common reverse primer. In yet another technique – ligase-mediated allele detection (Fig. 43.9D) – DNA ligase is used to couple two oligonucleotides at the site of a mutation. If one of the pair of nucleotides corresponds to either the normal or the mutant sequence, the ligase will only link the two oligonucleotides if there is a perfect match between the oligonucleotides and the target DNA, which has usually been amplified by PCR.

Because of the heterogeneity of genetic disease, one of the biggest obstacles to the application of DNA analysis is the requirement for simultaneous detection of multiple mutations. Since it is possible to perform several PCR reactions simultaneously ('multiplex PCR', Fig. 43.9E), several of the systems for the detection of mutations can be adapted to identify more than one mutation in a single assay. Alternatively, a reverse ASO method (in which oligonucleotides are immobilized and the test material is

used as a probe) allows simultaneous detection of several alleles. Such a principle underlies one form of the increasingly available 'microarray' or DNA chip, systems of DNA analysis. There are basically two forms of microarray. In the first, target DNA samples are immobilized on the chip and are then interrogated using labelled probes. The technique requires great precision in locating the sample sequences precisely on the chip, which needs to be done using robotic delivery systems. The second form is a reverse hybridization method in which the chip production procedure is highly complex but involves synthesizing thousands of different oligonucleotide sequences in situ on glass or silicon slides; these are the probes. The target DNA is amplified in a reaction that labels it with fluorophore, and allowed to hybridize with the immobilized probes. Bound label is detected using a laser scanner, with perfect sequence matches between sample and probe demonstrating stronger signals,

and the signals analysed using digital imaging software. The second method may be more appropriate to clinical analysis but the complexity of production means that chips tend to be produced by specialist manufacturers. At the moment, the instrumentation, and the chips themselves, remain costly but, as with all advances in molecular analysis, costs are likely to fall as the technology becomes more widespread.

In some inherited diseases, the gene responsible is particularly prone to deletions, copy number changes or duplications. Screening a gene by simultaneous amplification of several regions of DNA by multiplex PCR can be used effectively to screen such genes for deletions. Sequences that are absent or of decreased molecular weight can be detected simply by electrophoretic separation of the amplified sequences, although more precise methods have now been developed, for example multiplex ligation-dependent probe amplification (MLPA) (see Muscular dystrophy, p. 865).

Scanning or screening methods

To detect mutations when the precise site of mutation in a given individual is unknown, mutation screening methods were developed (Fig. 43.10). The sequence to be screened (e.g. one complete exon in a gene) is usually amplified first by PCR. One of several techniques that are capable of detecting the presence of single base differences in the sequence can then be used. In cleavage mismatch detection (Fig. 43.10A), hybridization of mutant and normal DNA strands produces a heteroduplex, and chemical or enzymatic techniques are used to cleave the strands at the site of mismatched base pairs. The size of fragments produced then allows localization of mutations within the sequence. Denaturing gradient gel electrophoresis and temperature gradient gel electrophoresis are methods that take advantage of the sudden decrease in electrophoretic mobility that occurs when a double-stranded molecule of DNA begins to dissociate (Fig. 43.10B). A homoduplex molecule (i.e. a duplex of complementary strands) will begin to dissociate and decrease its mobility at a characteristic point in a gradient of either denaturing agent or temperature. If denatured normal DNA is allowed to re-anneal in the presence of mutant DNA, heteroduplex molecules will form with a mismatch in almost complementary strands, which will begin to denature early so that the electrophoretic profile is altered. The technique of single-stranded conformational polymorphism (SSCP) is based on the fact that sequences of single-stranded DNA fold into specific conformations (Fig. 43.10C), so that normal and mutant DNA sequences can be separated by acrylamide gel electrophoresis. All of these, essentially manual, techniques are still used and remain valid techniques, but with rising workloads and the need for rapid results, they too are being increasingly overtaken by automated methods using large analytical platforms. Conformational sensitive capillary electrophoresis (CSCE) and denaturing high performance liquid chromatography (DHPLC) both utilize the principles of heteroduplex formation, with detection based on the different mobilities produced by conformational changes. Conformational

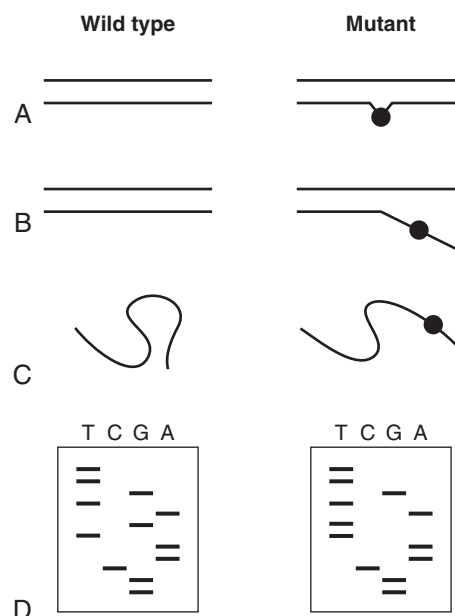


FIGURE 43.10 ■ Detection of unknown mutations. (A) Cleavage mismatch. If a heteroduplex is formed between wild-type and mutant DNA, unpaired bases at mutation sites are susceptible to chemical modification and subsequent cleavage. (B) Denaturing or temperature gradient electrophoresis. As DNA migrates through a gel with a gradient of temperature or denaturant, the rate of migration changes suddenly when the strands begin to separate. The presence of a mutation will alter the point at which this event occurs. (C) Single-strand conformational polymorphism. Molecules of single-stranded DNA form three-dimensional structures determined by their sequence, so that in non-denaturing gels, mobility is determined by sequence as well as length. The presence of a mutation may alter the three-dimensional structure formed and hence mobility. (D) DNA sequencing (dideoxy method of Sanger). Sequencing gels are read starting from the smallest fragment. Thus, the wild-type sequence is GGCAATGATGTT and the mutant sequence is GGCAATTATGTT.

sensitive capillary electrophoresis uses fluorescence detection to screen for the presence of mutations in a DNA sequence, while DHPLC uses altered retention times in an HPLC system.

If one of the above scanning techniques reveals the presence of a mutation, DNA sequencing can then be used to identify, or confirm, the precise mutation. Two techniques for sequencing of DNA were introduced in 1977, one by Sanger and the other by Maxam and Gilbert. Both techniques share the same basic principle in that a series of single-stranded DNA molecules, each one base longer than the last, is generated. These molecules can be separated by electrophoresis to generate a 'ladder' from which the sequence can be read. The dideoxy terminator method of Sanger (Fig. 43.10D) uses DNA polymerase to synthesize a complementary copy of the target DNA starting from a primer annealed close to the region of interest. The enzyme can incorporate dideoxynucleosides, but chain elongation then ceases immediately because these nucleoside analogues lack a 3' hydroxy group. DNA synthesis is carried out in four separate incubation mixtures, each containing the four deoxynucleoside triphosphate substrates (one of which is labelled) in addition to a low concentration of just one

of the four dideoxynucleoside analogues. When incubation is terminated, a population of labelled DNA molecules of varying lengths will have been produced. All molecules will have the same 5' end, but will vary in length to a base-specific 3' end (e.g. all terminating in A if dideoxy ATP was used). Originally, it was necessary to clone a fragment of DNA before sequencing could be carried out, but sequencing now uses PCR products obtained from either genomic DNA or cDNA. The use of radioactive (^{32}P) labelling followed by polyacrylamide electrophoresis (PAGE) gel and autoradiography has been superseded by DNA analysers which use four different dyes for the four different nucleosides, and automated read outs. The availability of these instruments, plus automated software for mutation and variant 'calling', makes sequencing more attractive for detecting mutations, and some laboratories use automated DNA analysers and sequencers as their front-line mutation detection system for some genes.

Tracking of mutant genes

When the gene causing a disease has not been identified, or it is not feasible to identify the precise mutation causing a defect, it may nevertheless be possible to predict, by gene tracking, whether a particular individual is affected provided a closely linked marker is available. Even when a disease gene has been identified and cloned, because the gene screening methods described above are not always 100% sensitive, gene tracking methods may still be useful to identify affected individuals. To be useful, a marker must exhibit a degree of polymorphism so that it is possible to distinguish between alleles associated with normal and mutant genes. Thus for a marker to be informative, an individual who is heterozygous for the disease locus must also be heterozygous for the marker. For genetic analysis, information on phase is also required, that is, it is necessary to determine which allele is linked to (tracks with) the mutation in an individual who is heterozygous. Some of the situations that may occur with a recessive disease are illustrated in Figure 43.11. Family 1 is fully informative and the disease gene is associated with allele a, so that the fetus would be affected only if it were also homozygous for allele a. In family 2, the analysis is informative only for the mother (in whom the mutant gene is linked to allele b), so that there is a 50% probability that the fetus can be predicted to be unaffected and a 50% chance that the fetus will be predicted to be at 50% risk. In family 3, the analysis is completely uninformative, since both parents are homozygous for the marker. In family 4, a marker with greater polymorphism is used, which is entirely informative.

Often a polymorphic marker can be found within the gene of interest, but if a linked marker outside the gene is used, there is an increased possibility of errors being made as a result of crossover. Even with intragenic markers, in large genes such as the dystrophin gene, there is a distinct possibility of crossover occurring between a marker and a mutation site. Similarly, there can be a degree of uncertainty as to whether an affected member of a pedigree has a new disease-causing mutation; again this may be particularly true in cases of

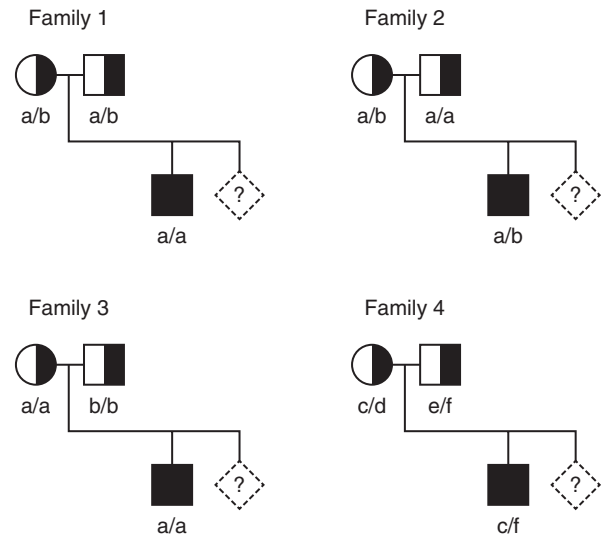


FIGURE 43.11 ■ Linkage analysis for a recessive condition. Linkage of alleles a, b, c, d, e and f is shown in four families. Half-shaded alleles symbols represent heterozygotes for the condition. Fully shaded symbols refer to affected individuals and diamonds refer to a fetus for which diagnosis is requested.

muscular dystrophy. The risk of incorrect diagnosis can be minimized by the use of multiple markers, but diagnosis will nevertheless be based on an estimate of probability. Other disadvantages of linkage studies are the expense, time and need to test other family members, so these techniques are used mainly in specialist laboratories. The developments of 'next generation' sequencing described below are likely to allow great improvements in identifying, and tracking, pathogenic mutations in such diseases and pedigrees.

There are several types of polymorphic marker that are suitable for tracking inheritance of particular alleles within families. Polymorphism at restriction sites (Fig. 43.12A) has been used extensively for linkage analysis, both for identification of the genes causing disease and for genetic analysis in families. Throughout the human genome, about one base in 1000 is polymorphic and around one in six random base changes creates or abolishes a restriction site. The majority of these polymorphisms are of no consequence to the organism, but their detection with restriction enzymes has provided what was, until recently, the most important technique for genetic mapping and linkage studies. Restriction fragment length polymorphisms (RFLPs) can be identified either by digestion of genomic DNA with the enzyme and subsequent use of a probe to identify fragments separated by electrophoresis or by PCR amplification of the region surrounding the restriction site, followed by restriction digestion and direct visualization of fragments after electrophoresis. A limitation of RFLP analysis is that the maximum polymorphism at any one site, can be no more than 50% (i.e. presence or absence of the site), and strategies that use relatively small numbers of SNPs have now been largely superseded.

However, the use of SNPs for gene tracking has now come full circle as a consequence of the Human Genome Project and the production of a dense SNP map of the

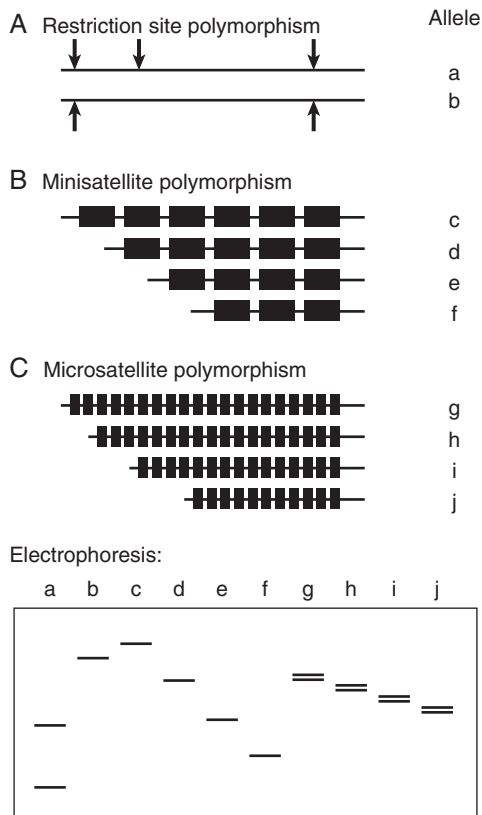


FIGURE 43.12 ■ Polymorphic markers used for linkage studies. (A) Alleles a and b are produced by polymorphism at a restriction site; (B) c, d, e and f, by polymorphism in the number of repeats in a minisatellite; (C) g, h, i and j, by polymorphism in the number of repeats in a microsatellite. Electrophoresis of a produces two small fragments after digestion, while b is not cleaved and remains as one large fragment. Electrophoresis of alleles c–j detects different sizes of mini- and microsatellites, but the latter may be complicated by separation of the two DNA strands (see text).

human genome. This knowledge, combined with the development of high-throughput microarray technology, means that it is possible to obtain SNP chips linked to large SNP databases; the use of large numbers of SNPs makes these chips more informative.

The discovery of hypervariable sequences in DNA consisting of a variable number of tandem repeats (VNTRs), with considerable polymorphism in the number of repeats, has found widespread application. Such sequences consist of short (10–60 base pairs) oligonucleotide sequences ('minisatellites') repeated 20–100 times, so that PCR amplification spanning the region (or excision of the region with a restriction enzyme) produces fragments which vary considerably in size in different individuals (Fig. 43.12B). The high degree of heterozygosity makes these VNTR loci particularly informative and, since they are inherited in a simple Mendelian fashion, they can be used in just the same way as RFLP analysis. Detection of these repeat sequences can be used, for example for determining zygosity in twins.

Similar repetitive sequences, with around 10–60 copies of motifs consisting of 1–4 bases (short tandem repeats or 'microsatellites') have been identified more recently (Fig. 43.12C). Unlike minisatellites, which are frequently

located near the ends of chromosomes, microsatellites are found scattered more evenly. One of the most common microsatellites, the (CA) n repeat (where n is roughly 15–30), occurs on average every 30 000 bases. Polymorphism in the length of these sequences can be detected by PCR amplification and used to track mutant genes in families. Dinucleotide repeats such as AC may separate into two bands on electrophoresis because the AC strand migrates faster than TG. Because fainter bands, which are probably generated during amplification, may also be apparent, there are advantages to using microsatellites with tri- and tetranucleotide repeats that are less susceptible to these problems, though the latter occur considerably less frequently in the genome.

Using a multilocus probe capable of hybridizing under low-stringency conditions to VNTR sites at several chromosomal sites, digestion with restriction enzymes gives a characteristic profile for each individual. The technique of 'DNA fingerprinting' has found most widespread use in forensic studies (e.g. the identification or elimination of crime suspects using samples of blood or semen) and in proving or disproving family relationships in civil cases (e.g. confirmation of maternity or paternity).

Analysis of these repeat sequences can now be performed quite straightforwardly by determining fragment length using capillary electrophoresis, a technique with greater accuracy than traditional gel based electrophoresis.

Next generation sequencing

One of the challenges in clinical genetics is the search for pathogenic mutations, where several genes may be implicated in a disease, or a group of disorders, for example, the peripheral neuropathies, and the clinical phenotype does not necessarily point to any one gene as the prime suspect, or where a large pedigree needs to be searched for potential disease causing mutations where one has not yet been identified in any individual of that pedigree. Traditional methods would involve testing one gene, or gene locus, after another. It would, however, potentially be more useful to screen all of the likely candidates, from all the members of a pedigree on a single sequencing run. The new high throughput technologies collectively called 'next generation' (or 'massively parallel' or 'deep') sequencing may well be the answer to this diagnostic problem. Several different platforms have been developed, with different sequencing technologies, but all the systems employ roughly the same essential steps.

- Step 1. Genomic DNA is fragmented using restriction enzymes.
- Step 2. The fragments are enriched for the regions of interest.
- Step 3. All the fragments are sequenced in parallel.
- Step 4. The results are aligned with a reference sequence to search for any variants (which at this stage may number in the thousands).
- Step 5. The bioinformatics step, comparing these variants to a polymorphism or SNP database (which may take the number down to the hundreds).
- Step 6. Variant calling, to determine if the variants detected are likely to be pathogenic or not.

Another common feature of these new platforms is that none of them use traditional Sanger type sequencing using dNTPs. Some of the techniques include:

- pyrosequencing, which uses luciferase to generate light when each of the individual nucleotides is added to the new DNA, with the reactions happening inside water droplets suspended in oil (emulsion PCR)
- use of reversible dye-terminators. Each molecule of DNA is amplified as a clone and fluorescently labelled nucleotides allow identification of each nucleotide as it is added before the dyes are removed prior to the next addition
- sequencing by ligation, whereby dye labelled oligonucleotides of known sequence are annealed and ligated to clones of amplified DNA immobilized on beads or glass slides
- ion-semiconductor sequencing, where the terminal incorporation of nucleotides uses traditional sequencing chemistry but detection of the incorporated nucleotides is achieved using a supersensitive ion sensor, essentially a form of pH meter, which detects the hydrogen ions released when each base is added, each cycle (taking only seconds) uses a different nucleotide but runs of the same nucleotide in the sequence will all be incorporated in the same cycle, giving a proportionally higher signal.

Even in the few years since the last edition of this book, both run (or 'read') times and cost per read have come down considerably, as has the size of the instruments, with recent ones very definitely being bench top, so why is it that these new techniques cannot automatically replace all the others currently in use? Over time, they may well do so, but at the moment the ability to sequence a whole genome or exome is not necessary for the detection of common pathogenic mutations in well characterized disorders, which tend to benefit from a targeted approach.

THE APPLICATIONS OF DNA ANALYSIS

The clinical application of DNA analysis has grown enormously since the first edition of this book, demonstrated by the increase, in both number and size, of molecular and clinical genetics departments in our hospitals.

Diagnosis of index cases

The very nature of inherited disease, with its implications for the families of affected individuals, means that most molecular testing now takes place in specialized genetics laboratories. These laboratories work closely with clinical geneticists and genetic counsellors. Consequently, any clinical biochemistry laboratory undertaking molecular testing must be aware of the consequences of a positive diagnosis. However, diagnosis of inherited metabolic diseases (e.g. amino acid and organic acid disorders) has, for many years, been undertaken in biochemistry laboratories, and the initial diagnosis of an inherited disease is still usually based on clinical history, physical signs and non-genetic laboratory tests. Confirmation of the diagnosis

can then often be obtained by a specific enzyme assay or protein study, but such assays frequently require the use of cell cultures, with their accompanying problems. Consequently, in an increasing number of diseases, DNA testing is now used for the diagnosis of index cases. These are often diseases in which (a) partial or complete gene deletions, which are relatively easy to detect, are frequent (as in muscular dystrophy) or (b) the defective protein has not yet been characterized or where it is difficult to devise methods for direct analysis of the protein (as with cystic fibrosis). One other potential advantage of DNA tests for diagnosis of index cases is that mutant genes are usually present in all nucleated cells, so that in diseases in which expression of the gene is limited to organs such as liver or kidney, biopsy can be avoided.

In the use of DNA analysis for diagnosis in index cases, it is important that accurate and reliable data exist on the relationship between genotype and phenotype.

Prenatal diagnosis

Analysis of fetal DNA, often with a view to selective termination of an affected pregnancy, can be used to determine the presence of severe or potentially fatal genetic disease. Such testing is most likely to be carried out when the parents have previously had an affected child, when they have been identified as carriers, or when one of them is affected by a dominant condition.

Although fetal blood samples can be used to detect many inherited diseases – particularly those that affect proteins in red blood cells or plasma proteins – DNA analysis can be carried out without taking tissue samples from the fetus. In the earliest prenatal diagnoses of genetic disorders, amniotic fluid cells obtained by amniocentesis at 16–18 weeks of gestation were used. However, it is now more common to use chorionic villus sampling (CVS), carried out at 8–10 weeks (see Chapter 22). Potential problems with prenatal diagnosis are contamination with maternal DNA and – especially where linkage studies are being carried out – uncertainty concerning paternity. Analysis of mini- or microsatellites can help avoid both of these problems. Where the diagnosis relates to an X-linked recessive disorder, if the fetus is shown to be female, either by cytogenetic analysis or a specific molecular test for identifying the Y chromosome, no further CVS testing may be necessary, as the fetus will, at most, be a carrier and this can be determined once the child is born.

Use of PCR makes testing of individual cells possible, so that it is possible to test fertilized embryos and implant only those known to be unaffected by a particular disease. This procedure has been carried out successfully with parents who are both heterozygotes for cystic fibrosis. The implications of this procedure are considerable, not least because many parents find selective abortion an unacceptable means of avoiding affected children. The use of such techniques for in vitro fertilization (IVF) is controlled, in the UK, by law, and is overseen by the Human Fertilisation and Embryology Authority.

A recent advance has been the ability to detect cell-free, fetal DNA in the blood of the mother, bringing with

it the possibility of testing the fetus for inherited disease without the need for invasive testing. Such testing does come with risks, mainly relating to avoiding contamination with the mother's own DNA. However, where the fetus is male, looking for markers carried on the Y chromosome only overcomes this problem. The use of DNA methylation ratios to separate fetal from maternal DNA has also been described. The use of fetal DNA in this way is very much in its early stages and is currently used primarily for determining fetal gender but is also being looked at as an alternative, and early, method of screening for Down syndrome.

Screening

The word 'screening' is used in various ways in medicine, and in the section on detection of mutations above, it was used in the context of examining a whole gene to find disease-causing mutations in patients with a clinical diagnosis of a genetic disease. In this section, the term is used in the context of testing healthy or presymptomatic individuals for molecular evidence of disease.

Screening of individuals

Screening can be 'targeted' in that individuals from families with strong family histories of a disease (high risk) are included and it may involve presymptomatic or carrier (heterozygote) testing. Such screening is often applied to determining the risks of developing a particular form of cancer, for example as in breast cancer screening. For women with several close relatives with the disease, there is a targeted screening programme based on DNA analysis. Although familial breast cancer represents a small percentage of cases of the disease, screening high-risk individuals is effective and involves examining the entire *BRCA1*, *BRCA2* and *TP53* genes.

Population screening

Population screening involves making available to all, on an equitable basis, testing for certain diseases or conditions. Examples of such programmes include antenatal screening for Down syndrome and those for cervical and breast cancer screening in women (not the same as the targeted testing described above). Currently, there are no such programmes that use DNA analysis in a primary testing strategy, although proposals have been made to use detection of fetal DNA in maternal serum for screening for Down syndrome, as mentioned above.

The criteria that must be satisfied for a successful screening programme are the same, regardless of whether traditional chemical analysis or DNA techniques are used (or, indeed, imaging or cytological testing) and have changed little in recent decades. All screening programmes are fraught with ethical and organizational problems, but none of these is peculiar to the analytical technique employed. The importance of education and the provision of counselling services for the success of a screening programme have been repeatedly emphasized, but the obtaining of informed consent, the reliability of methods (i.e. numbers of false positives and false

negatives), prevalence of disease in the population and the existence of a clearly recognized advantage to pre-symptomatic diagnosis (e.g. the possibility of avoiding precipitating factors in the environment) must be considered. In the UK, such screening has, since 1996, been overseen by the National Screening Committee, which also advises the government on the relevant issues.

Newborn screening to detect affected individuals with phenylketonuria and congenital hypothyroidism has been widely performed for some years using conventional biochemical tests on dried blood spots obtained by heel prick in the first two weeks of life. This scheme has now been extended to include testing for cystic fibrosis (see later section) haemoglobinopathies and medium chain acyl-CoA dehydrogenase deficiency (MCADD). While DNA diagnostic tests can also be carried out on blood spots, the programme still uses biochemical tests as the 'frontline' screening test. DNA testing can be used to follow-up positive screening results but currently is still employed mainly to identify the mutation(s) present in the child, rather than confirmation of the diagnosis.

At the present time, there are no plans to use DNA testing for primary screening of newborn infants. There are several reasons why this is not appropriate yet. In many genetic diseases, there can be large numbers of disease-causing mutations that produce the same biochemical (or haematological) effects. Similarly, not all mutations and variations in a gene will be pathogenic. Also, some mutations show variable penetrance or expressivity, so that whole gene screening, for example by sequencing, might create problems by identifying infants who have gene mutations but who will not necessarily develop the disease being screened for. Currently, it is still more efficient to screen by testing for the gene products. However, as DNA arrays and microchips become more accessible and cost-effective, molecular screening may become more widespread, providing there are accurate data on the genotype/phenotype correlations.

Pharmacogenetics

An area of laboratory testing in which biochemistry laboratories may find themselves involved is that of pharmacogenetics. Pharmacogenetics is concerned with identifying genetic variations that affect an individual's response to drugs. Initially, the term was used in respect of variations in drug metabolizing enzymes, but it has now been expanded to include polymorphisms in drug receptor, drug transporter and ion channel genes. Pharmacogenomics, on the other hand, refers to the (generally commercial) application of genomic technology in drug development and therapy. The two terms are often used interchangeably and there is some difference of opinion as to whether or not this is an acceptable practice. However, even though the differences may appear to be semantic, in general, pharmacogenomic studies in drug discovery regimes mainly aim to identify suitable drug targets and are concerned with multiple genes and phenotypes, whereas pharmacogenetics aims to identify variants in individual genes with the aim of 'personalizing' drug dosage and thus reducing adverse drug reactions. These are responsible for high levels of therapeutic

morbidity and mortality worldwide every year, with an associated cost to the health services of millions of dollars each year. Biochemical methods of identifying poor metabolizers (see below) are often cumbersome and demanding, using metabolic ratios that measure the relative concentrations of parent drug and metabolites in blood and urine.

The term 'pharmacogenetics', coined by Vogel in 1959, was originally brought to prominence by Kalow in descriptions of the variation in response to the muscle relaxant suxamethonium seen in patients with serum cholinesterase (butyrylcholinesterase) deficiency. Another of the earlier described pharmacogenetic effects was that observed with the enzyme N-acetyltransferase 2 (NAT2), where patients treated with the antituberculosis drug isoniazid were found to show variation in the rate at which they metabolized the drug: they were either slow, intermediate or rapid acetylators. The largest group of drug metabolizing enzymes is the cytochrome P450 (CYP P450) 'superfamily', which accounts for the metabolism of well in excess of 100 commonly used prescription and over-the-counter drugs. Of the large number of CYP P450 enzymes, six account for close to 90% of the drugs metabolized by this family. Of these, CYP3A4/5 account for around 50% but the most extensively characterized is CYP2D6, debrisoquine hydroxylase, accounting for 30%. It is the metabolizing enzyme for a large number and wide variety of drugs including antipsychotic, antiarrhythmic and antihypertensive drugs. Many drug metabolizing enzymes are now known to be highly polymorphic and many of them also demonstrate codominant inheritance, such that heterozygotes for normal and certain variant alleles show intermediate phenotypes. Poor metabolizers can be affected in a variety of ways, depending on whether it is the parent drug or its metabolite that is the pharmacologically active moiety. Thus, poor metabolizers may suffer the adverse effects of overdosing on a standard drug dose if the parent drug is active, or a poor therapeutic response if the metabolite is active. In the case of CYP2D6, more than 75 polymorphisms have been identified so far, and the prevalence of poor metabolizers and the frequency of different causative alleles vary significantly in different geographic and ethnic groups. Thus, in Caucasians, poor metabolizer status is found in 5–10% of individuals and is caused in nearly 90% of those cases by one of just three variant alleles, *CYP2D6*3*, *CYP2D6*4* and *CYP2D6*5*. (By convention, the commonest, or normal, allele of a gene is denoted as *1.) 'Extensive metabolizer' is the name given to individuals with an expected or normal response to a drug, and while these are mainly *1 homozygotes, some of the other alleles, such as the *2, also produce an extensive metabolizer phenotype. Intermediate metabolizers are thought to represent heterozygotes with extensive and poor metabolizer alleles.

The molecular variation responsible for the different alleles includes not only simple changes such as SNPs, small deletions and insertions but also whole gene deletions and duplications. The phenotypic effect of these differences can be on the concentration of enzyme produced or on its function, depending on the location of the polymorphism in the translated gene product. In the

case of CYP2D6, a duplication of the *2 allele leads to an ultra-rapid metabolizer phenotype.

Other clinically important drug metabolizing enzymes include thiopurine S-methyltransferase (TPMT) and uridine diphosphate glucuronosyltransferase 1A1 (UGT1A1). Thiopurine S-methyltransferase is involved in the metabolism of the immunosuppressant azathioprine and the antileukaemia drug 6-mercaptopurine; biochemical testing for TPMT phenotype has been available for some time but the role of genotyping has yet to be firmly established. Variants of UGT1A1 are responsible for Gilbert syndrome (unconjugated hyperbilirubinaemia) and Crigler–Najjar disease but the enzyme is also involved in the metabolism of the anticancer drug, irinotecan, and prospective genotyping may be useful in the future for avoiding some of the side-effects that can occur during treatment with this drug.

With the identification of the genes and sequences for an increasing number of drug metabolizing enzymes, analysis using the new techniques for rapid, high-throughput DNA testing should allow early identification of many of the variants, thus allowing altered drug dosage prior to, or shortly after, therapy commences, and so avoiding dangerous and costly adverse drug reactions.

Inherited diseases – some examples

Single gene disorders

The diseases most often thought of as 'genetic' tend to be those that arise from mutations in a single gene. Some disorders may arise from mutations in any one of a collection of genes related to a particular metabolic pathway, i.e. they may have monogenic causes but are not necessarily single gene diseases.

α_1 -Antitrypsin deficiency

The function of α_1 -antitrypsin (α_1 AT) is to inhibit neutrophil elastase and other proteases. The rate of association between α_1 AT and elastase is rapid, but after cleavage at the reactive site (methionine at position 358) of α_1 AT by the elastase, the molecule undergoes a radical change in structure that prevents dissociation of the elastase, thus inactivating it. The gene coding for α_1 AT (the protease inhibitor or PI gene) spans 12200 bases and is transcribed into a protein with a single chain of 394 amino acids.

α_1 -antitrypsin displays considerable polymorphism, with around 75 alleles having been detected so far. The variants are inherited in an autosomal codominant fashion, so that both alleles are expressed independently and can be detected in serum. Many of the variants, which are identified by letters of the alphabet depending on their electrophoretic mobility, function normally. The Z variant is of greatest clinical significance. In northern Europeans, around 5.3% of the population are carriers (i.e. are MZ heterozygotes) and about 1 in 2000 live births are ZZ homozygotes. Although deficiency can usually be detected by quantitation of serum α_1 AT, phenotyping by isoelectric focusing is more informative and reliable.

Secretion of α_1 AT, which is synthesized in the liver, is markedly reduced in ZZ homozygotes because the

mutant protein forms insoluble polymers, which accumulate in the endoplasmic reticulum of hepatocytes. The Z variant results from a point mutation converting a lysine residue in normal (M) protein to a glutamic acid one at position 342, which is at the base of the reactive centre loop. Juvenile cirrhosis occurs in 3–10% of ZZ homozygotes as a result of damage caused by the intracellular aggregates, though a much higher percentage are affected if a sibling has liver disease. The reasons for the incomplete penetrance of this condition are not known: intrauterine infection, gut-derived proteases, autoimmunity, fever and subclinical hepatitis have been suggested as possible exacerbating factors. Additional genetic factors, possibly in chaperone proteins responsible for directing misfolded protein to the secretory pathway, or other proteins involved in α_1 AT clearance have also been suggested.

Although other protease inhibitors are present in the lungs, α_1 AT contributes over 90% of the inhibitory activity capable of blocking neutrophil elastase. In the absence of α_1 AT, neutrophil elastase rapidly degrades tissue matrix components in the lung so that individuals with a markedly reduced plasma concentration of α_1 AT are prone to develop emphysema. As a result of reduced hepatic secretion of α_1 AT, the plasma concentration of this protein is reduced to about 15% of normal in ZZ individuals. Not all of these individuals are affected by disease, but 60–70% of ZZ homozygotes who smoke cigarettes develop pulmonary emphysema in the third or fourth decade of life. Smoking compromises the defences of the lung further because free radicals in cigarette smoke oxidize methionine at the reactive site of α_1 AT, drastically lowering its ability to inhibit neutrophil elastase.

Of the other mutations affecting the α_1 AT gene, the S mutation (changing the glutamic acid at position 264 to valine) is more common than Z, with an allele frequency of 2–4% in northern Europeans, but the consequences are less serious. Although some degradation of the S variant occurs in the liver and the serum α_1 AT concentration is reduced, S homozygotes are not considered to be at increased risk of liver or lung disease, although inheritance of the S allele with an allele causing severe deficiency (such as Z) confers a mild risk of emphysema. A number of other rare mutations (including the 'null' mutation) are associated with a decrease in plasma α_1 AT activity to a level that confers a risk of emphysema, and some others are characterized by accumulation of mutant protein in hepatocytes.

DNA-based techniques for identification of mutations in the PI gene complement phenotypic studies. Prenatal diagnosis is possible in juvenile cirrhosis families, where the causative mutation can be detected in the index case. DNA sequencing is the gold standard for detecting small deletions and point mutations, and this is readily available owing to the relatively small size of the gene. Various SNP detection techniques have also been developed for common mutations (e.g. Z and S), which have the potential to be used in screening programmes. While screening for this condition has been advocated, it has not been taken up widely owing to concerns regarding the variable penetrance of the condition, and the limited efficacy of interventions such as advice on smoking cessation.

Cystic fibrosis

Cystic fibrosis is inherited in an autosomal recessive manner. The homozygous condition is associated with defective exocrine secretion and consequent malabsorption with chronic obstructive pulmonary disease. The disease has a prevalence of about 1 in 2500 live births in northern Europe and a carrier frequency of 1 in 25, though it is much less common in other populations. For many years, laboratory diagnosis relied on the demonstration of raised chloride (and, optionally, sodium) concentrations in sweat. About 77% of affected individuals can be identified using the sweat test at two years of age and about 95% at 12 years. The test is far from ideal, because it is technically difficult to perform; not all patients with compatible clinical features have a raised sweat chloride and some individuals with a raised sweat chloride have no clinical features of cystic fibrosis. In one report, up to 40% of patients referred to cystic fibrosis centres had been wrongly diagnosed because of false positive or false negative sweat tests. Serum immunoreactive trypsin concentrations tend to be higher in affected newborns, forming the basis of a UK screening programme that can detect at least 95% of cases.

The gene affected in cystic fibrosis was mapped to 7q31 using linkage analysis. The gene itself was cloned in 1989 and named the cystic fibrosis transmembrane conductance regulator (CFTR). It has a size of approximately 250 kb with 27 exons and the CFTR protein is 1480 amino acids with a molecular weight of 168 kDa. The primary role of the CFTR protein is to form a chloride channel that reduces intracellular chloride. It appears to consist of two transmembrane domains, two ATP binding regions or nucleotide binding folds, and a regulatory domain. Phosphorylation of the regulatory domain by protein kinase A results in the opening of the chloride channel.

Over 1500 pathogenic mutations have been identified in the CFTR gene, although the great majority of them are extremely rare. The mutations include frameshift, nonsense, missense and splice site mutations and deletions. However, the commonest, and the first to be identified, is a 3-bp deletion in exon 10 of the CFTR gene that results in the loss of a phenylalanine codon at position 508. This mutation, Phe508del (Phe being the three letter code for phenylalanine and 'del' represents a deletion) represents around 70% of cases of cystic fibrosis. The frequency of this mutation in different populations varies markedly, however, and in Europe its frequency increases along a South East to North West gradient.

Detection of several cystic fibrosis mutations by multiplex PCR can be performed simultaneously using an ARMS assay (Fig. 43.13). In this assay, the common Phe508del mutation is detected together with 28 or 32 other mutations, depending on the kit manufacturer (e.g. the mutation causing substitution of Gly-551 by aspartic acid (Gly551Asp), another that converts Gly-542 to a stop codon (Gly542X) and a splice site mutation substituting T for G immediately after the last nucleotide in exon 4 (621+1G>T)), but chosen to be appropriate for the local population. This strategy will, for most

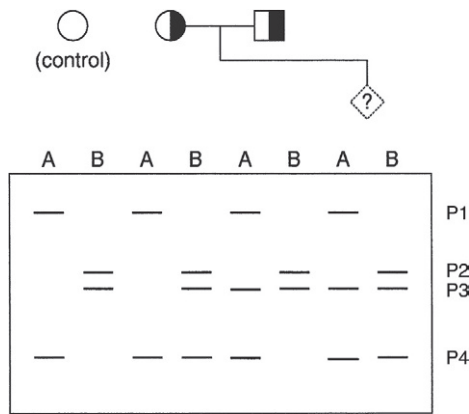


FIGURE 43.13 ■ Simultaneous detection of the common mutations in the cystic fibrosis gene by ARMS analysis. Bands P1, P2, P3 and P4 represent the product amplified using primers complementary to the sequences spanning the 621+1G>T, G551D, G542X and F508del mutation sites, respectively. Two reactions are carried out on DNA from each individual. In A lanes, P1 and P4 primers are complementary to the normal sequence, while P2 and P3 are complementary to the mutant sequences. In B lanes, P1 and P4 primers are complementary to mutant sequences while P2 and P3 primers are complementary to normal sequence. Thus, the mother is a carrier for the F508del mutation and the father is a carrier for the G542X mutation. The fetus has inherited both deleterious mutations (i.e. is a compound heterozygote). Adapted from Ferrie R M, Schwartz M J, Robertson N H et al. 1992 Development, multiplexing, and application of ARMS tests for common mutations in the CFTR gene. *American Journal of Human Genetics* 51: 251–262, with permission.

populations, identify the mutation responsible in more than 80% of cases. If the front-line test fails to reveal the mutation, additional mutations of progressively lower frequency are tested for and, ultimately, a full gene sequence may be the only way to identify a rare or new mutation.

Muscular dystrophy

Duchenne muscular dystrophy, one of the most common X-linked diseases (prevalence about 1 in 3000 live male births), results from mutations in the dystrophin gene and is characterized by progressive proximal muscle weakness in early childhood. Rarely, females can be affected, usually as a result of X-autosome translocations or skewed X inactivation. Symptoms are also apparent in 2–3% of female carriers ('manifesting carriers') caused by a non-random X inactivation. The diagnosis of this condition is often beyond doubt clinically, but confirmation is usually required because of its gravity. Plasma creatine kinase activity is markedly raised (often to 50–100 times normal), providing a useful confirmatory test. Becker muscular dystrophy, a less severe condition with onset in late childhood (affecting about 1 in 30000 newborn males), is an allelic disorder, that is, it also results from mutations in the dystrophin gene.

The dystrophin gene was discovered in 1986 and turned out to be an unusually large gene, currently the largest known gene in man, spread over around 2.5 million bases. The size of the gene may be one of the reasons why it is particularly prone to new mutations: it has been calculated that a third of all cases of

Becker muscular dystrophy arise from new mutations. Although at least 99% of the dystrophin gene is accounted for by introns, there are nevertheless 79 exons coding for a mRNA of 14000 bases. The dystrophin protein has a molecular weight of about 400000 kDa and it is expressed primarily in muscle cells. Dystrophin is normally localized to the inner surface of muscle cell sarcolemma and it is tightly linked with an oligomeric complex of glycoproteins that provides a linkage between the cytoskeleton and the extracellular matrix. When muscle biopsies of patients have been examined, it has been found that dystrophin is undetectable in patients with Duchenne muscular dystrophy, and present, altered in size or quantity, in patients with Becker muscular dystrophy.

In many patients (about 60%), the mutations causing disease are deletions or, less commonly (5%), duplications of exons. The remaining 30–35% are point mutations. Although the situation is undoubtedly more complex, in many cases it appears that Duchenne muscular dystrophy results from mutations that cause frameshifts, while the mutations causing Becker muscular dystrophy tend to be in-frame. Molecular diagnosis of muscular dystrophy has advanced considerably with the invention of PCR and its subsequent developments. Initially, diagnosis relied on RFLP linkage and cDNA probe analysis by Southern blotting. This technique was cumbersome and not always successful in identifying the disease mutation. An early use of PCR to detect deletions is shown in Figure 43.14, where ten pairs of primers were used to scan the whole dystrophin cDNA, allowing identification of some 65% of affected individuals. This was followed by the use of multiplex PCR, where deletions were detected by the failure of deleted exons to amplify in the reaction. Multiplex PCR has now been superseded as a front-line test by a technique called multiplex ligation-dependent probe amplification (MLPA). This technique involves the amplification of specifically hybridized

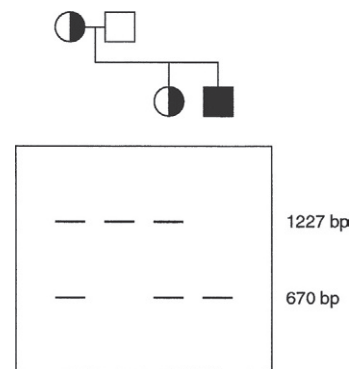


FIGURE 43.14 ■ PCR analysis to detect deletions in the dystrophin gene. Using PCR amplification to amplify the cDNA spanning exons 43–51, the mother is shown to be a heterozygote for the normal allele (band of 1227 bp) and a mutant allele (band of 670 bp) with a deletion of exons 45–48. The father is homozygous for the normal allele but the son expresses only the mutant allele. His sister, like their mother, is a heterozygote. Adapted from Roberts R G, Bentley D R, Barby TF et al. 1990 Direct diagnosis of carriers of Duchenne and Becker muscular dystrophy by amplification of lymphocyte RNA. *Lancet* 336: 1523–1526, with permission.

probes and enables the rapid identification of a deletion or duplication of any of the 79 exons in the dystrophin gene. It is particularly effective in determining the carrier status of females and has been used to identify the defect in families who appeared to be normal by earlier techniques. A series of characteristic 'printouts' is shown in Figure 43.15. It is also possible to detect point mismatch. In families in which the defect in one allele is not due to deletion, either multiplex PCR or RFLP analysis can be used for prenatal diagnosis. Multiplex PCR can also be carried out on dried blood spots for neonatal screening. However, in the absence of any effective treatment, this will not benefit the child being screened. Any advantage conferred by the opportunity for counseling to avoid further affected children must be balanced against the effect on the parents of having to explain to their son that he has a fatal illness, and, indeed, the effect of this knowledge on the child himself.

Huntington disease

In Huntington disease, progressive chorea (irregular involuntary movements) and dementia lead to death on average 17 years after the onset of symptoms. The condition is autosomal dominant with a prevalence in the UK of 3–7 per 100 000. The trait shows complete penetrance, but age-dependent expressivity: only 10% of affected individuals have symptoms by 30 years, but this proportion reaches 95% by 70 years. Thus, most individuals who are at risk reach the age at which they may wish to have children without knowing whether or not they will develop the disease.

As a result of exhaustive research by several groups, the genetic defect causing Huntington disease was first located on chromosome 4; the gene (named *IT15*) has since

been pinpointed to 4p16. Inheritance of many disorders, including fragile X syndrome, myotonic dystrophy and spino-bulbar-muscular atrophy (Kennedy disease), is linked to a mechanism in which increasing expansion of triplet repeats eventually leads to disease. Huntington disease is caused by this type of mutation, with expansion of a CAG repeat in exon 1. The variable age of onset and severity of the disease correlates with the extent to which the CAG repeat has been expanded. Normal alleles contain up to 26 repeats, are stable and not associated with disease, while disease genes contain more than 39 repeats. Those with 27–35 repeats have the potential either to decrease or to expand and thus become disease alleles: they are 'mutable'. Alleles with 36–38 repeats show reduced penetrance with some heterozygotes never developing symptoms. Disease-associated alleles containing more than 39 repeats are prone to large increases in repeats from one generation to the next, which results in a phenomenon called anticipation, where symptoms develop at a younger age.

Before the identification of the mutation, presymptomatic diagnosis using RFLP analysis and linkage studies was performed with a probability that depended on the number and informativeness of family members who could be tested. Not only did this technique leave a degree of uncertainty over the diagnosis, but diagnosis of one individual might depend on analysis of samples from other members of the family who did not wish to be tested. This technique is now used only rarely and the diagnosis is now made using PCR analysis to measure directly the size of the CAG repeat region.

The ability to identify patients with Huntington disease many years before symptoms become apparent raises some of the ethical problems of DNA testing in

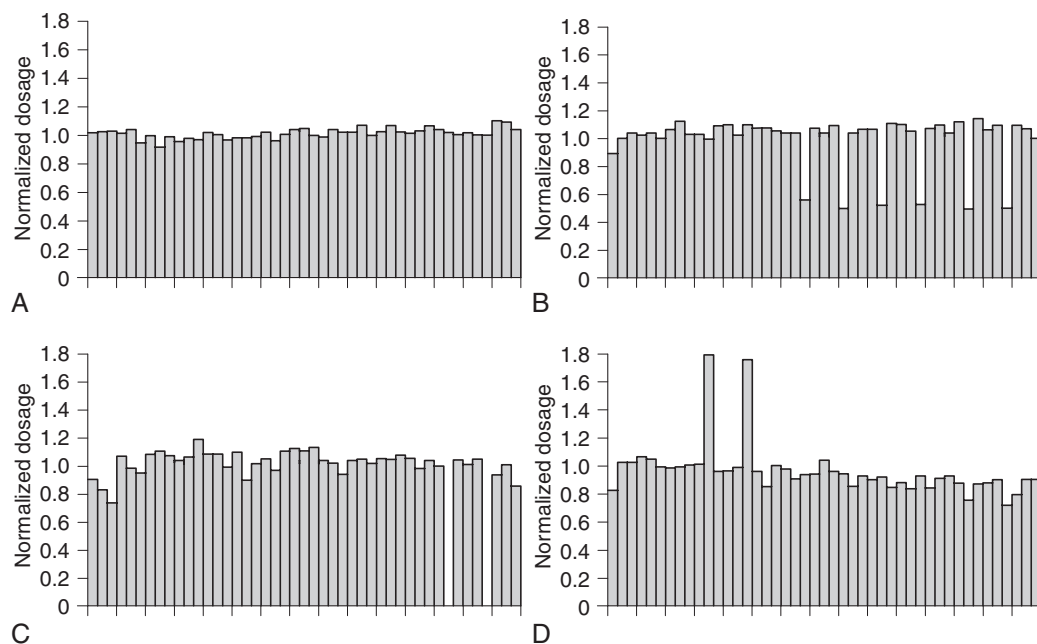


FIGURE 43.15 ■ Results of MLPA analysis of the dystrophin gene showing normalized ratios of the hybridization signal: (A) is from a normal individual; (B) shows deleted exons in a female carrier; (C) shows deleted exons in an affected male and (D) shows duplicated exons in an affected male. In (B) (C) and (D) the affected exons are contiguous in the gene but not in the readout. Courtesy of the Bristol Genetics Laboratory, North Bristol NHS Trust.

their most acute form. Potential psychological problems for individuals without symptoms knowing that they have a fatal disease must be considered and appropriate counselling provided. The risk of stigmatization, with difficulties in employment and obtaining insurance, makes absolute confidentiality of results crucial. The fact that the disease is transmitted in a dominant fashion also raises issues where members of one generation wish to be tested but those of another do not, as a positive result in a child almost certainly implies a positive result in the parent (depending on the repeat number). Consequently, in the UK at least, predictive testing is only undertaken using a robust protocol involving close working between clinical genetics and laboratory genetics teams.

Multifactorial and polygenic disease

While much progress has been made in understanding single gene disorders, and developing techniques for their diagnosis, such diseases are still rare in most populations. The current challenge lies in understanding those common diseases that have a polygenic or a multifactorial basis. Where a disease or disorder arises from mutations in a small number of genes, each of which contributes, it can be said to be polygenic. Diseases where the role of genetic factors is combined with environmental triggers can be thought of as multifactorial. However, in reality, conditions such as diabetes and essential hypertension, which affect millions of people worldwide, can be both and remain the focus of a lot of current investigation.

Atherosclerosis

Atherosclerosis is one such multifactorial disease and the recognition of a correlation between plasma cholesterol concentration and coronary artery disease has provided the background for genetic studies of several candidate genes. (A search for atherosclerosis on the Online Mendelian Inheritance in Man (OMIM) website gives more than 200 'hits'.) Environmental factors such as diet are clearly important for determining plasma cholesterol concentrations, but it has been estimated that at least 50% of the population variance of cholesterol concentration has a genetic basis, with only a small group being monogenic (requiring no, or virtually no, environmental triggers for their expression). Transport of lipids between tissues is a complex process involving formation, modification and clearance of lipoprotein particles (see Chapter 37). The plasma concentration of one of these particles, low density lipoprotein (LDL), is positively correlated with the risk of atherosclerosis. LDL particles consist predominantly of cholesteryl esters and an apolipoprotein (apoB), and their removal from plasma depends on binding of apoB to a specific cell surface receptor (the LDL receptor). Abnormalities in both the apoB and the LDL receptor genes have been shown to affect clearance of LDL, both resulting in elevated plasma cholesterol concentrations with greater elevations caused by abnormalities of the LDL receptor than of apoB.

Familial hypercholesterolaemia

Familial hypercholesterolaemia (FH) is a dominant disorder affecting about 1 in 500 individuals and which accounts for around 5% of individuals with clinically evident atherosclerosis. In the UK, the most recent guidance from the National Institute for Care and Excellence (NICE), Clinical Guideline 71, and the subsequent Diagnostics Guidance published in 2011, recommends that a diagnosis of FH should be made based on the Simon Broome criteria, which include DNA-based as well as clinical/biochemistry criteria, and that those with a clinical diagnosis should be offered a DNA test. Of the mutations associated with FH approximately 93% are in the LDL receptor gene (*LDLR*). It is, however, a heterogeneous condition and mutations in the genes for apoB, proprotein convertase subtilisin/kexin type 9 (*PCSK9*) and LDL receptor-associated protein (*LDLRAP*) are now known also to cause it (see Chapter 37).

The LDL receptor gene is located on chromosome 19, spans 45 kilobases and codes for a protein of 839 amino acids. More than 400 mutations, many of which are deletions, have been identified. These mutations can be divided into five main classes:

1. 'null' alleles that produce no protein
2. mutations that block transport of the newly synthesized receptor from the endoplasmic reticulum to the Golgi apparatus
3. mutations that prevent the binding of LDL at the cell membrane
4. mutations producing receptors that can bind but cannot internalize LDL
5. mutations producing receptors that cannot release LDL following internalization and so are not recycled back to the cell surface – receptors must be freshly synthesized.

The apoB gene (*APOB*) is located at chromosome 2p24 and codes for a protein of 4536 amino acids. Apolipoprotein B-100, the full-size protein that is found in LDL, is produced by the liver while a smaller protein, apo B-48, is produced in the intestine by a unique tissue-specific process that introduces a stop codon into the mRNA. Familial defective apoB-100 is a dominant disorder, with a frequency of around 1 in 800 in the general population. Unlike deficiency in the LDL receptor, it is a much more homogeneous condition. Most cases result from one mutation, a SNP affecting codon 3500, which decreases the affinity of apo B-100 for the LDL receptor. Mutations in *APOB* account for around 1.5% of FH cases.

The third gene for which the NICE Diagnostics Guidance now also recommends testing is *PCSK9*, which encodes proprotein convertase subtilisin/kexin type 9. This protein is involved in degradation of the LDL receptor. A gain of function mutation of *PCSK9* results in a reduction in the number of LDL receptors at the cell surface, thus reducing LDL uptake into cells and leading to increased circulating cholesterol. There remains some uncertainty as to the percentage of cases attributable to *PCSK9* mutations.

FH, arising from mutations in the three genes described above is inherited in autosomal dominant fashion

but a very small percentage (<1%) show an autosomal recessive mode of inheritance. Mutations in the *LDLRAP* gene (previously referred to as the autosomal recessive hypercholesterolaemia gene, *ARH*) are one of the causes of this rarer form of LH. Patients with ARH have a normal LDL receptor but the receptor associated protein, encoded by *LDLRAP*, fails to mediate internalization of the receptor in the usual way.

Apolipoprotein E genotypes

Apolipoprotein E (apo E) mediates clearance of two other classes of lipoprotein particle (chylomicron remnants and intermediate density lipoprotein, IDL) by binding to the remnant receptor (IDL can also bind to the LDL receptor). There are three common apoE isoforms, E2, E3 and E4, characterized by different electrophoretic mobilities, with frequencies of 10, 30 and 60% respectively. The E4 variant is associated with increased cholesterol concentrations in comparison with those associated with E3, while E2 is associated with lower concentrations. Familial dysbetalipoproteinaemia (remnant, or type III, hyperlipidaemia) is associated with the E2/E2 phenotype, but although almost all individuals with the condition are E2 homozygotes, the genotype shows poor penetrance so that most E2 homozygotes do not manifest the disorder, and other factors in addition to the E2 allele are clearly involved in expression of disease.

It has been estimated that about 7% of the variance in plasma cholesterol LDL concentrations found in populations can be accounted for by recognized mutations in the apo B, LDL receptor and apo E genes. Other candidate genes that may eventually prove to be involved in the complex process of atherosclerosis include those involved in cholesterol absorption, intracellular cholesterol metabolism, haemostasis and fibrinolysis. Angiotensin-converting-enzyme (ACE) regulates the concentrations of factors (angiotensin II and bradykinin) that are involved in modulation of vascular tone and proliferation of smooth muscle cells, and some recent results suggest that a polymorphism in this candidate gene is a risk factor for atherosclerosis. Two polymorphic forms of ACE have been described called the I (insertion) and D (deleted) forms, arising from the presence or absence of a 287bp sequence in the gene. Individuals homozygous for the D form have been shown to be at increased risk of atherosclerosis, particularly those considered to be at low risk from other criteria.

Cancer genetics

Cancer can be considered to be a genetic disease. This is because it arises either from somatic mutations in cells that then become cancerous or from inherited germline mutations that lead to a predisposition to the disease, although cancers arising from a single gene defect account for fewer than 5% of cases. The genetic component is often triggered by environmental or behavioural factors (such as cigarette smoking, which increases the risk of developing lung cancer) and it is now clear that cancer is the end-point of an accumulation of somatic and germline

mutations, most notably in proto-oncogenes and tumour suppressor genes.

Oncogenes and suppressor genes

The oncogenes were first identified as genes in retroviruses capable of producing tumours in birds and rodents. The first of these genes was identified in 1973, when a single gene (*src*) of the Rous sarcoma virus was shown to be capable of producing sarcomas in chickens. Later, it was shown that precursors to oncogenes, the proto-oncogenes or cellular oncogenes, were present in normal cells and that viral oncogenes were copies of these normal genes that had become incorporated into the retroviral genome. These viral oncogenes are able to transform infected cells once activated in some way.

Although it is now known that viruses do not cause most of the common cancers in humans and most of the oncogenes detected in this way (including *src*) are not of prime importance in human cancer, work with viral oncogenes made possible the rapid advances in understanding of human tumours that are now taking place. The isolation of human oncogenes was facilitated by the development of transfection techniques, in which DNA isolated from tumours can be tested for ability to transform cultured cells to a cancerous phenotype. Mechanisms for activation of proto-oncogenes include:

- chromosomal translocations so that the oncogene comes under the influence of regulatory elements from other genes
- amplification of a region of DNA including the oncogene
- point mutations that confer constitutive activity on the gene product.

Because the activation of only one allele is sufficient for transforming activity, oncogene mutations are generally dominant.

More than 50 cellular oncogenes have now been identified; the proteins they code for are mostly involved at some stage in the cascade of events associated with stimulation of cell division by growth factors. Thus, the *sis* protein is a mutated form of the platelet derived growth factor (PDGF); *erbB* codes for a truncated form of the epidermal growth factor (EGF) receptor; *erbA* codes for a mutant form of the thyroid hormone receptor; *ras* (a GTP-binding protein related to the hormone receptor G proteins) is a transducer of growth factor responses, and the *jun* and *fos* oncogenes mediate growth factor-induced gene expression. Oncogenes such as *mos* have been linked to factors that control the cell cycle.

In normal cells, the growth-promoting effects of proto-oncogenes are thought to be balanced by growth-constraining suppressor genes (tumour suppressor genes). Early evidence for the existence of such genes came from experiments in which fusion of tumour cells with normal cells resulted in hybrids with properties of normal cells. These experiments were interpreted as evidence for a tumour-suppressing gene in normal cells whose activity had been lost in tumour cells. Retinoblastoma, a rare form of eye tumour occurring in children, was later shown to result from loss of both copies of a gene (*RB*) located in

the q14 band of chromosome 13. In familial retinoblastoma (around 40% of cases), a defect in one *RB* allele is inherited, so that tumours arise in cells in which the remaining (normal) allele is lost. In the remaining 60% of cases (sporadic cases), both alleles in the tumour founder cell have undergone somatic mutation. This is the two-hit hypothesis of Knudson, who proposed that the disease followed two-hit kinetics but that, in familial cases, the first 'hit' is inherited, that is, is a germline mutation, and the second is somatic.

Most cancers do not follow the retinoblastoma paradigm in such a straightforward manner, but one that does is neurofibromatosis type 1, associated with tumours derived from the embryonic neural crest and caused by a defect in the *NF-1* gene located in the 17q11.2 chromosomal region. As with retinoblastoma, predisposition can be inherited from an affected parent as a germline mutation and tumour formation is initiated in cells that lose the activity of the remaining normal allele. The effects of suppressor gene mutations are usually only apparent when both alleles are inactivated, so that most of these genes behave in a recessive manner.

The majority of tumours that lack a functional suppressor gene (such as *NF-1* or *RB*) are found to have two identical mutant alleles. Elimination of the normal alleles is thought to occur in these cases by mechanisms such as chromosomal non-disjunction or gene conversion, which replace the normal allele with a copy of the mutant allele at high frequency (10^{-3} – 10^{-4} per cell generation). By looking for evidence of this process, suppressor genes have been identified through demonstrating 'loss of heterozygosity' at specific chromosomal sites. Anonymous, highly polymorphic, DNA markers that identify heterozygous sites in normal tissue were used to demonstrate reduction to homozygosity in tumours, indicating loss of one or more alleles (Fig. 43.16), although this approach is not without pitfalls. However, the development of DNA arrays has been useful in demonstrating the possible location of tumour suppressor genes specific for different tumour types against the general background of non-specific ones. One pitfall is that simple loss of heterozygosity is not the only molecular abnormality in tumours; many demonstrate multiple structural abnormalities and some of the observed losses may be due to deficient DNA repair mechanisms or to chromosomal instability rather than the selective loss of a tumour suppressor gene. One likely mechanism for inactivation of tumour suppressor genes involves methylation of specific CpG dinucleotides in tumour suppressor gene promoters. In some cases, methylation may be an alternative to point mutation, whereas in others it appears to be the only mechanism for loss of function.

One of the most studied tumour suppressor genes is *TP53* (which encodes the transcription factor p53), the loss of which is a major contributor to genomic instability and is possibly the commonest single genetic change seen in cancer. The function of p53 appears to be related to apoptosis (programmed cell death) so that – unlike normal cells – any cell with a defective p53 protein is more likely to proliferate despite the acquisition of mutations. Defects in this gene, which is located at 17p12, can be from mutation or deletion, and p53 can be eliminated by inhibitive action of other gene products such as that of *MDM2*.

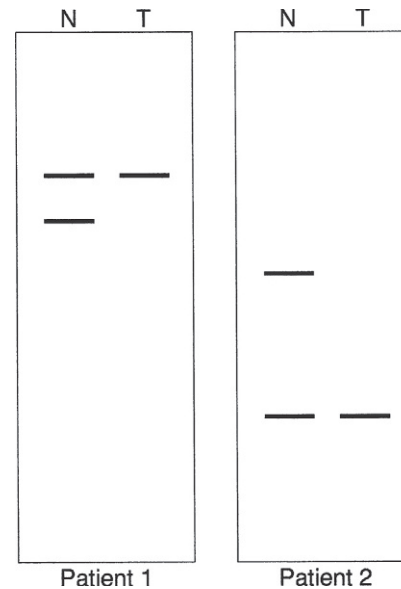


FIGURE 43.16 ■ Loss of heterozygosity in tumours. Electrophoretic separation of digested total genomic DNA and detection of an allele on chromosome 5 with a single-locus probe in normal tissue (N) and tumour (T) of two patients with colorectal tumours. In both cases, the probe detects the alleles on DNA fragments of different size in normal tissue, demonstrating heterozygosity. Loss of heterozygosity in tumour tissue is demonstrated by disappearance of one allele. Adapted from Solomon E, Voss R, Hall V et al. 1987 Chromosome 5 allele loss in human colorectal carcinomas. *Nature* 328: 616–619, with permission.

Inherited mutations of *TP53* are found in Li–Fraumeni syndrome, a dominantly inherited condition in which sufferers demonstrate multiple primary tumours, and are also implicated in some familial breast and colon cancers.

Another mechanism whereby suppressor genes function is illustrated by the 'deleted in colonic carcinoma' (*DCC*) gene, identified initially by loss of heterozygosity in the long arm of chromosome 18, which occurs in over 70% of colonic carcinomas. Sequencing of this gene showed that it encodes a 190 kDa protein with properties that suggest that it might be a transmembrane molecule that binds cells to the extracellular matrix or basement membranes.

The protein encoded by *RB* (pRB) has been shown to form a complex with the oncoproteins produced by the adenovirus SV40 and the human papillomavirus, suggesting that the ability of these viruses to form tumours results from their ability to inactivate *RB*. During the cell cycle, pRB switches from hyperphosphorylated to relatively unphosphorylated forms, so that pRB is likely to be involved in regulating the cell cycle.

In solid tumours, full malignancy requires acquisition of the ability to metastasize. The processes involved include changes in adhesion molecules, proteases and angiogenic factors. Little is known about the genetic changes that activate these processes, but it may be a side-effect of the overall genomic disarray seen in the cells of advanced tumours.

It has been recognized for many years that tumour cells become increasingly genetically unstable, accumulating mutations more rapidly as the tumours grow. Genetic changes, which may be acquired or inherited, can play

a critical role in tumour progression by influencing mutation rates rather than growth regulation. Genes with this type of function are well illustrated by xeroderma pigmentosum, a rare autosomal recessive disease caused by defects in nucleotide excision DNA repair. Sufferers are extremely sensitive to ultraviolet light so that skin exposed to sunlight develops large numbers of freckles, which often progress to skin cancer as the damage caused by the sun cannot be repaired.

The early emphasis of research on genetic changes in cancer has been on understanding the mechanisms by which oncogenes and suppressor gene abnormalities can bring about transformation of a normal cell. The knowledge obtained thus far is now leading to progress in pre-symptomatic detection of tumours, and of identification of high-risk individuals, as well as in their diagnosis and treatment. This is demonstrated by the breast cancer screening programme for women at high risk of developing the familial form of disease. Sequence screening of the entire *BRCA1* and *BRCA2* genes in high-risk individuals (around 5% of cases) has close to 100% sensitivity, thus allowing prophylactic mastectomy when individuals are shown to have inherited the abnormal gene but, significantly, permitting reassurance for those women who have not. Testing regimens for other forms of cancer, for example bladder cancer, are being developed based on the fact that tumours shed cells. Sensitive PCR techniques can amplify the DNA of tumour cells shed into urine, blood and faeces, which can then be examined for SNPs (possibly using high-density arrays), loss of heterozygosity and microsatellite instability in characteristic tumour suppressor and oncogenes. Although not in routine use, such systems, as well as being useful for early detection of tumours, are likely to be valuable for monitoring possible disease recurrence.

The focus of clinical cancer genetics is now not only development of screening programmes capable of detecting tumours at earlier stages, but on the determination of whether or not the tumours are malignant, and how responsive they are likely to be to treatment. Studies on colorectal tumours have shown that formation of malignant tumours requires mutations in at least four or five genes, but that fewer changes are present in benign tumours. Work on glioblastoma tumours has shown that the degree of methylation in the O⁶-methylguanine DNA methyltransferase (MGMT) gene can predict response to chemotherapy. O⁶-methylguanine DNA methyltransferase is a DNA repair enzyme that removes toxic alkyl groups from the O⁶ position of guanine. Epigenetic silencing of MGMT by promoter CpG methylation has been associated with longer overall survival in patients with glioblastoma who, in addition to radiotherapy, received alkylating chemotherapy with an alkylating agent such as temozolomide. High levels of MGMT activity in cancer cells create a resistant phenotype by reducing the efficacy of alkylating chemotherapy.

Multiple endocrine neoplasia (MEN)

Not all 'cancers' or neoplasms are malignant, and some have both benign and malignant forms, a fact demonstrated by the MEN group of disorders (see Chapter 41).

As the name suggests, these disorders are characterized by tumours of endocrine glands; they are familial and much is now known about their genetics. MEN1, characterized by tumours of the pituitary, parathyroids and pancreas, is inherited in a dominant fashion with a high penetrance. The gene involved is located on chromosome 11, at 11q13, and codes for the protein menin. The exact function of the protein remains unknown but at the time of writing is thought likely to be a tumour suppressor gene, as the majority of mutations found in MEN1 patients appear to result in loss of function. Pathogenesis follows Knudson's 'two-hit' hypothesis as described above, whereby tumour development occurs in individuals who have inherited the first 'hit' as a germline mutation in the *MEN1* gene when a somatic mutation occurs in a relevant endocrine cell. MEN1 is notable in that many different mutations have been described as contributing to the condition.

MEN2 also demonstrates autosomal dominant inheritance and consists of three sub-groups, all of which have medullary carcinoma of the thyroid in common. In the case of MEN2, the gene responsible is the *RET* proto-oncogene on chromosome 10, which codes for a receptor tyrosine kinase. However, unlike in MEN1, the germline *RET* mutations of MEN2 result in gain of function effects and the site of the mutation appears to be critical in determining the site of the tumour. There are also fewer loci within the gene where the small number of activating mutations are likely to be found, of great value in the molecular diagnosis of MEN2 and in contrast to the very many in the menin gene leading to MEN1.

GENE THERAPY

Gene therapy can be characterized in several ways, but essentially, it is the use of introduced genetic material to correct disease, either to replace a defective gene product or to correct an abnormal gene. It should not be confused with the treatment of genetic disease, which currently still uses conventional therapies. Ethically, correction of genetic defects by insertion of genes into somatic cells is akin to organ transplantation. However, germline gene therapy, where introduced genes could be transmitted to future generations, is universally agreed to be unethical for use in humans, although the production of transgenic animals by introduction of genes into fertilized eggs is widely used for studying gene function and regulation. Consequently, many countries have regulatory bodies to oversee gene therapy programmes, which concentrate on somatic cell gene therapy. Such programmes target the cells, organs or tissue affected by the disease or disorder under investigation. The first replacement of a defective gene in a human (to correct immune deficiency resulting from adenosine deaminase deficiency) took place in 1990, with limited success in that while none of the ten patients in the trial was cured, no adverse effects were reported. Since that first trial, gene therapy has had a difficult history, but the picture has improved since the last edition of this book, with nearly 2000 clinical trials approved in the last five years. As the field is once again changing rapidly, this section is intended to provide an outline of the issues and principles involved.

To be a suitable candidate for potential gene therapy, it is an essential prerequisite that the gene involved has been cloned and sequenced, together with all the appropriate promoter and regulatory elements. The next step requires a mechanism for introduction of the therapeutic DNA into suitable target cells, and the affected tissue or organ must be identified and accessible. The introduction of the DNA into the target needs a vector, which may be viral (such as retroviruses or adenoviruses) or non-viral (such as liposomes) and can be *ex vivo*, where the patient's own cells are cultured with the vector and then reintroduced, or *in vivo*, where the transformed vector is delivered directly to the affected tissue or organ. *Ex vivo* techniques are generally to be preferred as the cells can be checked before they are returned to ensure that the desired change has been achieved.

With retrovirus vectors, the viral genome is integrated into the DNA of infected cells after reverse transcription of viral RNA into DNA. Disadvantages in the use of retroviruses are, first, the relatively small amount of DNA that can be introduced (less than ~7 kb), and second, that they are unable to infect non-dividing cells, which limits their use as very few cell types are continually dividing. Adenoviruses avoid these problems, since they can carry much larger segments of DNA and are also able to infect non-dividing cells, but they do have the problem of potentially being immunogenic. Indeed, a crucial part of the development of a gene therapy strategy is that the vector should produce no harmful effects. One potential danger of gene therapy using a viral vector is the initiation of cancer if the inserted gene disrupts the function of a cellular oncogene or suppressor gene. This is called insertional mutagenesis and such an effect was seen in trials for treating the immunodeficiency disorder X-linked severe combined immunodeficiency (XL-SCID), using *ex vivo* enrichment of patients' lymphocytes with a retroviral gene vector. This trial was initially seen as a great success, with 9 of the 11 patients being cured, but two of them later developed leukaemia owing to an insertional activation of the oncogene *LMO2*. This adverse outcome led to the suspension of gene therapy trials using pools of lymphocytes. Such a problem occurs mainly with the use of retroviruses and can be avoided by using adenoviruses, but they do have the problem of immunogenicity, and are known to contain genes of their own that can be involved in the process of malignant transformation and thus may also induce malignancy but by a different mechanism. It has not been possible, thus far, to determine precisely where in the host genome the introduced gene will insert: the process is random. In the case of the XL-SCID trial, only a small number of insertions were in a position to activate the oncogene but they produced clones that outgrew the other, beneficial, clones. Consequently, until, or unless, it is possible to direct the insertion site, or avoid damaging insertions, protocols using random insertion of vectors are unlikely to be approved.

A further problem is that the introduced normal gene must be capable of expressing its normal product in sufficient amounts to be effective in correcting the host disorder. Surprisingly, it has been found that functional genes can be transfected to skeletal muscle by direct injection, but this approach does not appear to be successful

for other tissues. This approach has had some success in treating haemophilia where adeno-associated virus expressing factor VIII was injected intramuscularly. The encouraging results of this strategy are thought to be because only a small increase in the amount of factor VIII has a major clinical benefit.

Non-viral methods of gene therapy include the use of 'naked' DNA, that is, the direct injection of DNA into target cells, and the use of liposomes, aqueous vesicles with an outer lipid bilayer, as vectors for the foreign DNA. Non-viral methods have the advantage of being non-immunogenic and so of being safer. The use of directly injected DNA is likely to be of use mainly where a small amount of product will have a significant benefit, as in the example above of factor VIII. Liposomes have the advantage of being able to carry larger quantities of DNA, but the expression of the introduced gene is short lived, meaning that repeated treatment is necessary. A potentially useful development of liposome-mediated transfer has been in the use of DNA-protein complexes designed to target cell surface receptors. The use of an appropriate protein, recognized by the cell receptor, leads to internalization to intracellular vesicles of the complex. This allows transport to lysosomes where the complex is degraded and, once the gene has escaped from the lysosome, it can be expressed.

Stem cells in gene therapy

Correction of defects in haemopoietic cells is of particular interest because stem cells in bone marrow are relatively accessible and diseases that can be corrected by bone marrow transplantation are obvious candidates for gene therapy, bearing in mind the difficulties associated with random insertion vectors described above. This is most likely to be feasible in the correction of enzyme deficiencies such as Lesch-Nyhan and Gaucher diseases, where effective therapy could be achieved with relatively low levels of gene expression.

Early stem cell work looked at the use of embryonic cells. However, such work raised ethical concerns and more recent studies have looked at inducing pluripotency in somatic cells (called iPS cells), such as fibroblasts, by introducing genes known to be involved in maintaining pluripotency. A hope for these iPS cells is that they could be used in autologous therapy in patients with genetic disease once the gene defect had been corrected in the iPS cells *in vitro*. A major target for gene therapy has been cystic fibrosis. In this disorder, some of the most damaging effects result from a lack of the gene protein in the lungs, and studies have indicated that restoration of as little as 5–10% of normal gene expression would be enough to produce a beneficial clinical response. An adenovirus, tropic for respiratory epithelium, was successfully exploited as a vector to insert these genes while other trials used liposomes or adeno-associated virus. However, although gene transfer was demonstrated, there were concerns about the safety of using adenoviruses that might cause an immune reaction and in both types of trial the gene expression was too short-lived. If stem cells, such as iPS cells, could be used, this could avoid some of the need for repeat administration and thus reduce the risks of an immune reaction.

Early trials using stem cells as gene therapy in this way, to correct some rare forms of inherited eye disorders, have shown some success and hold promise for the future of these techniques.

Gene therapy in cancer

A major aim of gene therapy has been the treatment of cancer, with more than 60% of approved trials targeting this area. Projects are looking at various stages of the process, including the supplementation of tumour suppressor genes and the prevention of activated oncogene expression, but also at areas such as the manipulation of tumour cells to promote apoptosis or to render them vulnerable to the host's immune system by increasing their antigenicity. Such trials demonstrate the diverse potential of gene therapy.

So, while the work carried out in the field of gene therapy has mushroomed in the last decade, there are still many problems to be overcome before its use in clinical medicine becomes frequent or widespread. However, the knowledge gained in the process has been enormous and it may take only one significant breakthrough to produce a major leap forward.

CONCLUSION

Clinical molecular genetics has advanced tremendously in recent years, a fact demonstrated by the almost exponential rise in the establishment of genetics laboratories in hospitals in developed countries. Its perceived importance in the future of healthcare in the UK was demonstrated by the publication, in 2003, of a Department of Health White Paper, *Our Inheritance, Our Future*, which resulted in large injections of money into NHS genetics laboratories to fund investment in new technology and staffing and was followed by a review document in 2011.

The advances of the past ten years have perhaps not been in the direction that might have been predicted at the start but some of the most startling have been in technology. Surprisingly, some techniques, such as DNA chips and microarrays, have been slower to transfer to routine clinical use than might have been expected, almost certainly because the cost of the analytical platforms and consumables did not fall as quickly as hoped. However, other technologies such as CSE and real-time PCR have become routine and the speed with which next generation sequencing is being applied to clinical genetics is remarkable. Consequently, predictions of high-throughput, faster turnaround analysis have become a reality – and technological development is still enabling scientific discovery.

Gene therapy is beginning once again to live up to its early promise with more successes being reported and safer solutions coming forward. Conversely, pharmacogenetics has grown much more slowly than predicted, despite evidence that shows that predictive genotyping can prevent adverse drug reactions, the reasons for this remain unclear, but unexpected complexity of the issues may play a part.

The emphasis in the coming years is likely to be 'more and faster'; that is, increasing laboratory capacity for molecular analysis and improving turnaround times. In terms of innovation, areas such as direct analysis of cell-free DNA in plasma may well begin to make an appearance in the laboratory.

The role of molecular biology in polygenic and multifactorial diseases is receiving a lot of attention, especially as these are the conditions (diabetes, hypertension, heart disease) that affect a large proportion of the population. Linked to this is the growing area of epigenetics, an area that examines heritable characters that are not due to changes in the DNA sequence. One mechanism for such inheritance is methylation.

Other areas that are finding their way into the clinical laboratory now are expression analysis and proteomics. At its simplest, expression analysis (sometimes called 'transcriptomics') examines the products of genes by creating cDNA from a gene's mRNA (usually by reverse transcriptase PCR – rtPCR) and cloning it into vectors from which the protein product can then be expressed and studied. Expression screening employs microarray technology to examine heterogeneous mixtures, usually of mRNA or cDNA, often to compare expression in different tissues. An area worthy of a book in its own right, proteomics takes expression analysis one step further, in that it is concerned with the analysis of complex mixtures of proteins, that is, with the expressed mRNAs that are actually translated. Not all mRNAs are translated into protein products and some undergo post-translational changes. Proteomics allows identification of changes in the relative abundance of expressed proteins as well as of their structure and function using mass spectroscopy techniques, in particular matrix-assisted laser desorption-time of flight mass spectrometry (MALDI-TOF MS). The technique is expected to be particularly valuable not only in the study of cancer (e.g. bladder cancer), but also, it is hoped, in its diagnosis and treatment.

Clinical molecular biology has made huge advances in recent years but there are still many more to be made.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the invaluable help of the staff of the Bristol Genetics Laboratory, North Bristol NHS Trust and of David Halsall, Department of Clinical Biochemistry, Addenbrooke's Hospital, Cambridge in the preparation of the earlier edition of this chapter. The chapter is based, with permission, on that written by Dr Michael Norman, Department of Medicine, University of Bristol, for the first edition of this book.

GLOSSARY

- Allele** alternative forms of a gene at the same locus
Balanced polymorphism a polymorphism that is stable in a population
Carrier an individual who is heterozygote for a recessive gene

cDNA DNA that is complementary to a mRNA molecule

Clone a cell line derived from a single cell, or gene sequences propagated by recombinant DNA techniques

Codominant pertaining to two alleles which are both expressed in a heterozygote

Compound heterozygote an individual with two different mutant alleles at the same locus

Crossover (or recombination) exchange of information between homologous chromosomes during meiosis

Diploid chromosome complement with two copies of each chromosome (as in normal human cells where the diploid number is 46)

Dominant an allele that manifests its phenotypic effect in heterozygotes

Downstream sequences further in the direction of expression (5' → 3')

Enhancer a regulatory DNA sequence that can function to stimulate transcription of a gene irrespective of its position and orientation relative to that gene

Epigenetic heritable traits (e.g. those caused by gene methylation) that are not dependent on DNA sequence changes

Exome all the exons in a genome

Exon any segment of an interrupted gene that is present in the mature mRNA

Expressivity variability in the severity of a genetic trait

Frameshift mutation mutation resulting from insertion or deletion of bases (but not a multiple of three) that alters the reading frame of an mRNA

Gamete haploid cell generated by meiosis (sperm or egg)

Genome the complete ensemble of genetic information of an individual

Germ cell gametes or their precursors

Haploid cell a cell containing only one copy of each chromosome

Haplotype a group of closely linked alleles that are inherited as a single unit

Heterozygote an individual with different alleles (usually one normal and one mutant) at a given locus on homologous chromosomes

Homozygote an individual having the same allele at a given locus on homologous chromosomes

Hot spot site with a high frequency of mutation or recombination

Imprinting differences in the expression of genes depending on parental origin

Intron any segment of an interrupted gene that is transcribed but removed during formation of mature mRNA

Linkage two or more loci on a single chromosome that are sufficiently close that they do not segregate independently in offspring

Linkage disequilibrium association of two loci more frequently than predicted by chance

Locus unique location on a chromosome of a gene or particular DNA sequence

Loss of heterozygosity (LOH) homozygosity (in a tumour or somatic cell) when the constitutional state is heterozygous

Meiosis a series of two modified mitoses generating haploid gametes from a diploid cell

Messenger RNA (mRNA) sequence of RNA transcribed from a gene which, after processing, codes for a protein

Microsatellite polymorphic sequences due to a variable number of tandem repeats of a di-, tri- or tetranucleotide sequence

Minisatellite polymorphic sequences due to a variable number of tandem repeats of a short sequence of ten or more nucleotides

Mitosis process of division in somatic cells

Multiplex PCR simultaneous PCR reactions performed using more than one pair of primers in the same reaction mix

Mutation a heritable change in the genetic material

Northern blot a technique for transferring RNA to a filter for subsequent detection

Oncogene a gene involved in cell development capable of causing transformation to a tumour cell

Polymerase chain reaction (PCR) a technique for amplifying a specific sequence of DNA

Penetrance the frequency with which a particular genotype is expressed

Phenotype the observable characteristics of an individual

Point mutation changes in the sequence of DNA involving single base pairs

Polygenic a trait influenced by the cumulative effects of several genes at different loci

Polymorphism the occurrence of two or more alleles at a given locus at significant frequencies in the population

Positional cloning cloning of a gene after determining its chromosomal position by linkage analysis without knowing its function.

Probe a labelled fragment of DNA used to identify complementary sequences by hybridization

Promoter region of DNA to which RNA polymerase binds before initiating transcription

Recessive an allele that produces a phenotypic effect only when present in the homozygous state

Restriction enzyme an enzyme that cleaves DNA at specific sequences

Retrovirus an RNA virus that utilizes reverse transcriptase to insert itself into the DNA of a host cell

Restriction fragment length polymorphism (RFLP) polymorphism in the size of fragments produced by digesting DNA with a restriction enzyme

Somatic cell all cells of the body except gametes

Southern blot the technique of transferring fragments of DNA after electrophoresis to a filter

Trait any phenotypically detectable character or property

Transcription production of mRNA from the DNA template

Transfection incorporation of foreign DNA into a cell

Transformation conversion of cells to a state of unrestrained growth, resembling tumour cells

Transgenic animal animal into which a foreign gene has been incorporated

Translation conversion of the mature mRNA message into a protein

Tumour suppressor gene a gene that is growth constraining and whose inactivation can lead to unrestrained growth of a cell

Upstream sequences located in the opposite direction to transcription (3' → 5')

Vector any plasmid, phage etc. into which foreign DNA can be inserted for cloning

Variable number of tandem repeats (VNTR) minisatellite and microsatellite sequences with polymorphism in the number of repeats

Western blot transfer of proteins to a filter after electrophoresis

Wild type the allele which is most frequent in natural populations; now referred to as 'normal type'.

Further reading

In a subject that changes as rapidly as this one, any bibliography is likely to become outdated very quickly and so this one is short with respect to texts, listing only those likely to form a valuable 'core' for the subject. Readers are directed to the internet, where the most up-to-date information is to be found: several appropriate starting points are listed. Regular review of appropriate journals, many of which are available online, is also recommended, especially for technological applications.

Background

Strachen T, Read PR. *Human molecular genetics*. 4th ed. London: Garland Science; 2011.

Essential text covering the principles of molecular genetics; well referenced.
Turnpenny P, Ellard S. *Emery's elements of medical genetics*. 14th ed. London: Elsevier; 2012.

Highly readable text for clinical aspects, covering cytogenetic as well as molecular genetic disease with a useful chapter on ethical issues.
Young ID. *Medical genetics*. Oxford: Oxford University Press; 2005.

Although slightly older now, this is another very readable introductory text that uses landmark publications and case histories to illustrate the issues effectively.

Internet resources

The British Society for Human Genetics. <http://www.bshg.org.uk>.

An excellent place to start: it is a searchable site and has extensive links to other major websites as well as to useful online genetics journals.

OMIM database, linked through the NCBI website (PubMed). <http://www.ncbi.nlm.nih.gov/omim>.

For information on Mendelian disorders, the OMIM site contains highly referenced entries for most disorders and genes. The NCBI website is also the starting point for links to other databases, e.g. on protein structure, as well as being the best starting point for literature searching.

Journals

Clinical Chemistry. The Journal of the American Association of Clinical Chemistry (AACC).

Now subtitled as an International Journal of Molecular Diagnostics and Laboratory Medicine, this is an excellent source of up-to-date information on emerging molecular techniques and their applications, which also tend to be of particular relevance to clinical biochemists.

Nature Genetics.

The dedicated genetics journal in the 'Nature' family of scientific journals.

Forensic biochemistry

Robert J. Flanagan • Sarah Belsey • Terhi Launiainen

CHAPTER OUTLINE

INTRODUCTION 874

SAMPLES AND SAMPLING 875

POISONING WITH ENDOGENOUS AGENTS 876

γ-Hydroxybutyrate 876

Insulin 877

Magnesium 877

Sodium 878

POST-MORTEM BIOCHEMISTRY 878

Vitreous humour 879

SPECIFIC DIAGNOSTIC PROBLEMS 880

Anaphylaxis/anaphylactoid reactions 880

Diabetes 880

Drowning 880

Hypothermia/hyperthermia 880

Inflammation 882

Sudden death 882

INTRODUCTION

Forensic biochemistry can be defined as the application of biochemical assays in the service of the courts. Examples include DNA analysis for human identification and methods for the detection of trace evidence, such as the Kastle–Mayer (phenolphthalein/hydrogen peroxide) and luminol reactions used to detect the presence of blood. This chapter, however, focuses on laboratory measurements rather than methods. Many of these are standard laboratory procedures, while others are specific to forensic work.

The range of forensic situations that the clinical biochemistry laboratory may be asked to help with is wide and, as in all such work, the results of tests can usually be properly interpreted only when considered together with all the available evidence. Post-mortem biochemistry has a role in investigating the cause of death in some apparently natural deaths, including both diabetic and alcoholic ketoacidosis, deaths that may have involved a prolonged stress response such as hypothermia, as well as in the diagnosis of disease processes such as early myocardial infarction, which may be difficult to diagnose by physical examination. There is clearly considerable overlap between clinical and forensic toxicology, in that some endogenous substances can be used as poisons (e.g. sodium chloride, potassium chloride and insulin) and in some instances, suspicion of poisoning may be aroused by abnormal biochemical results (see [Table 44.1](#)).

Forensic biochemistry has an important role in the investigation of deaths and serious injuries occurring in hospital. Assault committed within hospital may involve poisoning, and can range from murder (intent to kill), through manslaughter (culpable homicide) and attempted murder, to malicious poisoning (usually of a child or an elderly relative). Iatrogenic poisoning may range from

relatively minor drug administration errors, to catastrophes such as asphyxia caused by an anaesthetic error. The results of an analysis, or residual or unused samples, or even apparatus used in giving the drug, may be required by officers acting on behalf of the courts. Therefore, in all cases, careful specimen labelling and handling, reporting and laboratory record-keeping is important. The standard of completion of request forms and sample labelling is still very poor in some hospitals. This can cause many problems if samples are required by the police or coroner. Knowledge of the limitations of the analytical methods used is also important – an enzymatic ethanol assay is not as selective as headspace gas chromatography, for example.

One practical problem is that samples may be collected, assays performed and results reported before the need for forensic investigation has become apparent. Another problem is that all that may be available are specimens collected after death (see [Box 44.1](#)), although in general it is information on an analyte's concentration prior to, or at the time of death that is required. In this case, the likelihood of agonal or post-mortem change, and indeed, sample contamination during collection, must be taken into account when interpreting results. An associated problem is that there are often no reference ranges for fluids such as vitreous humour, pericardial fluid, or synovial fluid since such samples are, for practical purposes, rarely available during life, except from laboratory animals. Method validation is also compromised by this same lack of reference material. Furthermore, the time needed for analyte equilibration between plasma and, for example, vitreous humour during life remains unknown.

Of course the laboratory, too, can be the subject of forensic investigation if laboratory error of whatever nature comes under scrutiny of the courts (see [Box 44.2](#)). Examples here include the use of inappropriate analyses,

TABLE 44.1 Some laboratory investigations commonly requested on blood samples that may arouse suspicion of poisoning

Investigation	Possible cause of increase	Possible cause of decrease
Sodium	MDMA ^a (malignant hyperthermia), sodium salts	Diuretics, water intoxication (acute and chronic), MDMA (very rare)
Potassium	Digoxin, potassium salts	Diuretics, laxatives (both chronic), insulin, salbutamol, sulfonylureas, theophylline
Glucose	Salicylates, theophylline	Ethanol (especially children), insulin, salicylates, sulfonylureas, valproate
Calcium	–	Ethylene glycol, fluorides, magnesium salts
Chloride	Bromide or organobromines (actually interference in method)	–
Lactate	Ethylene glycol (artefact on some blood gas analyzers)	–
Magnesium	Magnesium salts	–
International normalized ratio (INR, prothrombin time)	Anticoagulant rodenticides (warfarin, brodifacoum), paracetamol (early marker of hepatic damage)	–
Anion gap ($[(\text{Na}^+) + (\text{K}^+)] - [(\text{HCO}_3^-) + (\text{Cl}^-)]$)	Ethanol, ethylene glycol, iron salts, isoniazid, methanol, metformin, paraldehyde, salicylates, toluene (chronic)	–
Osmolar gap ^b	Acetone, ethanol, ethylene glycol, methanol, 2-propanol, hypertonic i.v. solutions (e.g. mannitol)	–

^aMethylenedioxymetamphetamine.

^bMeasured osmolality (freezing point depression) – calculated osmolality. Calculated osmolality = $2([\text{Na}^+] + [\text{K}^+]) + \text{urea} + \text{glucose}$ (all mmol/L).

delayed analyses, sample mix-ups and errors in reporting, including the use of inappropriate units. In all cases, it is best to write out units in full, for example ‘milligrams per litre’, rather than using symbols, when producing reports for the courts. The laboratory should provide clear guidance as to the significance of a result, especially when different units may be used. Patients have died when a paracetamol result reported in mg/L has been assumed to be in mmol/L. Mass units (SI) should be used for drugs except for lithium, thyroxine and methotrexate, where molar units should be employed. For metals/trace elements and for alcohol (ethanol), either molar or mass units may be employed. However, for forensic purposes, including the regulations governing occupational lead exposure, mass units are often the rule and the laboratory should remember this when providing interpretation of results. For clinical purposes, ethanol is often reported as mass units per litre (mg/L), but for forensic purposes, at least in the UK, ethanol is still reported as mg/100 mL (mg%).

SAMPLES AND SAMPLING

Information recorded on the sample container at the time the sample is collected should include the names (first and family or last name) and date of birth, patient or post-mortem number and the date and time of collection. This information, together with details of the collection site in post-mortem work, and the sample type (including a note of any preservative), and any other appropriate information, should be recorded on an accompanying assay request form (see Box 44.3). The date and time of receipt of all specimens by the laboratory should be recorded and a unique identifying number assigned in each

BOX 44.1 Thanatochemistry

Post-mortem biochemistry is sometimes termed ‘thanatochemistry’ (from the Greek, *Thanatos*, the personification of death in Greek mythology).

BOX 44.2 Red or dead?

A 64-year-old male with a long history of medical problems, including type 2 diabetes, was admitted to hospital at approximately 17:00 with a 2-day history of diarrhoea. A blood sample was requested on admission for electrolytes and assessment of renal function. However, phlebotomy was difficult; blood was obtained at the sixth attempt (22:12). The results were sodium 128 mmol/L, potassium 8.2 mmol/L, urea 36.8 mmol/L and creatinine 596 μmol/L. The results were known in the laboratory at 23:40, but were not reported because the sample was haemolysed. Instead, the laboratory asked for an urgent repeat sample.

In the event, no attempt was made to obtain a second specimen urgently in view of the difficulty in obtaining the first sample. The patient was found collapsed at about 01:10 the next day. Resuscitation was unsuccessful. The high urea and creatinine indicated that the patient had been in renal failure for some hours before he died and, if the urea and creatinine results had been reported together with the potassium, this would have alerted the clinician to the possibility of a potentially fatal hyperkalaemia and placed the likelihood of haemolysis contributing to the raised potassium in proper clinical context.

In retrospect, it was agreed that appropriate action to lower the plasma potassium should have been taken. Insulin, nebulized salbutamol and/or haemodialysis (depending on the clinical condition of the patient and the facilities available) would have been the treatments of choice.

BOX 44.3 Information that should accompany a request for forensic biochemistry or toxicology

- Name, address and telephone number of clinician/pathologist and/or coroner's officer, and address to which the report and invoice are to be sent. A post-mortem (reference) number may also be appropriate
- Circumstances of incident (including copy of sudden death report if available)
- Past medical history, including current or recent prescription medication, and details of whether the patient suffered from any serious potentially infectious disease such as hepatitis or tuberculosis
- Information on the likely cause, and estimated time, of ingestion and/or death and the nature and quantity of any substance(s) implicated
- If the patient has been treated in hospital, a summary of the relevant hospital notes should be supplied, including details of emergency treatment and drugs given both therapeutically and incidentally during investigative procedures
- Note of occupation/hobbies
- A copy of any preliminary pathology report, if available

case. Any residual specimen should be kept securely at -20°C or below until investigation of the incident has been concluded.

In forensic work, it is important to be able to guarantee the identity and integrity of the specimen from when it was collected through to the reporting of the results, although this ideal is often not attained in normal clinical laboratory practice. 'Chain of custody' is a term used to refer to the process used to maintain and document the history of the specimen (see Box 44.4). Procedures for appropriate storage of samples that may be required for forensic analysis and for authorizing and documenting the release of such samples on request of a coroner, for example, must also be in place. Ideally, samples should be protected during transport by the use of tamper-evident seals and should be submitted in person to the laboratory by the coroner's officer or other investigating personnel. If storage is to be at -5 to -70°C , basic precautions to preserve sample integrity, including labelling, must be undertaken (glass tubes will break if over-full when frozen). The requirements of the UK Human Tissue Act, or

BOX 44.4 Chain of custody documents

- Name of the individual collecting the specimen
- Name of each person or entity subsequently having custody of the specimen, and details of how it has been stored
- Date and time the specimen was collected or transferred
- Specimen or post-mortem number
- Name and date of birth of the subject or deceased
- Brief description of the specimen
- Record of the condition of tamper-evident seals

other relevant legislation on the retention and storage of pathological samples, must be met.

POISONING WITH ENDOGENOUS AGENTS

Many analytes of toxicological interest (and their metabolites) also occur naturally in the body and hence, reference ranges, 'cut-offs' or other means of delineating an exogenous source for the compound(s) of interest must be adopted. Examples include acetone, carbon monoxide (measured as % carboxyhaemoglobin) ethanol, γ -hydroxybutyrate (GHB), insulin, iron, potassium, sodium and testosterone.

Testosterone continues to be abused by athletes and this remains the most common adverse finding declared by World Anti-Doping Agency accredited laboratories; the practice is usually detected by the demonstration of an elevated testosterone:epitestosterone ratio and by procedures to ensure that such a finding is not 'natural' for a particular athlete. Numerous other endogenous agents have been used in an attempt to enhance performance or to mask use of other agents in sport but further discussion of this topic is beyond the scope of this chapter.

γ -Hydroxybutyrate

γ -Hydroxybutyrate (GHB), and its precursors γ -butyrolactone (GBL) and 1,4-butanediol, are used to improve athletic performance, as intoxicants and sometimes in assault such as drug-facilitated sexual assault (which includes rape). As with alcohol, voluntary ingestion always has to be considered (Box 44.5). γ -Hydroxybutyrate exposure may also arise from unexpected sources (Box 44.6). The substance is practically odourless and tasteless and has a short plasma half-life (20 min or so), making covert administration hard to detect. Quantitation is needed as GHB is an endogenous compound. A urine cut-off of 10 mg/L is recommended in attempting to differentiate endogenous GHB excretion from deliberate GHB/GBL administration. Unusually, the urine cut-off seems more reliable than a measure of plasma GHB, but by ~ 12 h after GHB ingestion, the concentration in urine is usually <10 mg/L. For post-mortem blood, the cut-off is

BOX 44.5 Covert GHB administration?

A man and a woman met at a party. They spent the evening together and took a taxi to the woman's house around midnight. She had no further recollection of events until she awoke and thought that she had had sexual activity. She reported the suspected incident to police and provided blood and urine samples (fluoride preserved) about 20 h after the incident. Cannabinoids and GHB (30 mg/L) were present in urine. No GHB was detected in either blood (limit of detection 5 mg/L) or drinks containers (glasses etc.) at her home.

The man was arrested and charged with rape on the day the incident was reported. A bottle of 'Lucozade' in his possession contained about 20% (w/v) GHB. However, although he was found guilty of rape, the woman later admitted to prior voluntary GHB ingestion.

BOX 44.6 Girl nearly killed by 'GHB' toy

A girl of 7 years nearly died after eating some toy beads. She suffered severe poisoning. The beads were coated with 1,4-butanediol, a GHB precursor. The girl said she ate the beads thinking they were sweets: they tasted of marzipan. Six children in Australia and NZ and two in the USA needed hospital treatment after swallowing them. The beads, made in China, could be arranged into designs and fused together when sprayed with water. Half a million Bindeez toys were recalled in the UK/Eire in November 2007.

generally taken as 50 mg/L because of the likelihood of post-mortem GHB production.

Insulin

Endogenous insulin is co-secreted with C-peptide. In theory, therefore, in suspected administration of exogenous insulin, which lacks C-peptide, the ratio of the two analytes could be useful as an indicator of the source of the insulin. However, commercially available immunoassays have highly variable responses to different types of insulin making the insulin:C-peptide ratio extremely difficult to interpret. An insulin-degrading enzyme (IDE) is widely distributed in tissues including erythrocytes, hence use of haemolysed samples, including post-mortem samples, is likely to give false results. Use of different anticoagulants to preserve samples is also associated with differences in plasma insulin results. The presence of hepatic or renal insufficiency, or of endogenous anti-insulin or anti-proinsulin antibodies (analogous to the analytical problems with digoxin assay if anti-digoxin Fab antibody fragments have been given) may be further possible sources of error (see also Chapter 17).

Plasma C-peptide is stable for only ~2–3 weeks at -20°C and for up to 6 months at -80°C , whereas plasma insulin is more stable (~5 h at room temperature, ~1 week at 4°C and several months at -20°C). On the other hand, C-peptide is not degraded by IDE. Anti-insulin antibodies bind proinsulin via its insulin moiety and greatly retard its clearance from the circulation. Because of cross-reaction with proinsulin in some C-peptide immunoassays, proinsulin bound to anti-(pro)insulin antibodies can interfere.

Murder or suicide by insulin is difficult to diagnose and to prove (see Box 44.7). Immunohistochemical demonstration or measurement of elevated insulin concentration in tissue around an injection site compared with a control site can help support the diagnosis. Unlike most purely immunological methods, liquid chromatography-tandem mass spectrometry (LC-MS/MS) may allow differentiation of human insulin from its synthetic derivatives. Besides its use in doping control, the method has been applied to post-mortem material related to an insulin poisoning case.

Magnesium

Poisoning with magnesium is rare, and usually follows intravenous administration. Confusion over units is not uncommon (see Box 44.8). Clinical features of

BOX 44.7 The Grantham disaster

February to April 1991: three children died suddenly on Ward 4 at Grantham and Kesteven Hospital. A further baby died at home soon after discharge. Nine other patients collapsed unexpectedly, some more than once. All but one were transferred urgently to the paediatric ITU. Clinical and post-mortem investigations were inconclusive. One patient had three hypoglycaemic attacks on Ward 4, and insulin poisoning was suspected.

On 12 April, a very high insulin/low C-peptide result was reported on a blood sample collected on 28 March. On 30 April, police were called to investigate continuing emergencies (which included a further death). In November 1991, enrolled nurse Beverly Allitt, 26, was charged with four murders, nine attempted murders and nine counts of causing grievous bodily harm. She was also charged with the attempted murder of two adult patients when she had worked on their ward before moving to the children's ward. Potassium and lidocaine were implicated in some of the Ward 4 assaults.

In 1993, she was convicted of four murders, three attempted murders and on six counts of causing grievous bodily harm. She was found not guilty with respect to the charges concerning the adult patients. As in all criminal poisoning cases, much of the evidence was circumstantial. Allitt's history suggested factitious disorder (Münchhausen syndrome).

BOX 44.8 Was it poisoning?

A 57-year-old male presented to hospital complaining of chest pain and dizziness. He was sweaty and light-headed (heart rate reported as 230/min). Ventricular tachycardia was diagnosed, and three unsuccessful attempts were made at electrical cardioversion. His condition deteriorated.

Cardiology advice at this stage was to give 8 mmol magnesium ion by slow i.v. bolus followed by further magnesium by infusion. In the event, an i.v. bolus of 8 g magnesium sulphate (32 mmol magnesium ion) was administered, whereupon the patient complained of feeling flushed and unwell. He was visibly hot and sweaty, and suffered cardiac arrest. After 13 cycles of cardiopulmonary resuscitation and administration of calcium chloride to reverse possible magnesium toxicity, he was pronounced dead some 2 h after presentation.

Whole blood magnesium was 1.50 mmol/L in a sample obtained on admission and was 2.67–3.54 mmol/L in three samples obtained after the administration of magnesium sulphate, but before death. These latter results provided evidence to support the report of accidental magnesium overdose, but the blood magnesium concentrations are similar to those obtained in patients given magnesium to control convulsions.

At post-mortem the next day, femoral whole blood magnesium was 10.0 mmol/L and vitreous humour magnesium (each eye separately) was 0.46 and 1.42 mmol/L. In court, it was suggested that:

1. the post-mortem blood magnesium result was likely to have been affected by post-mortem change
2. the average of the vitreous humour results (0.94 mmol/L) was compatible with the ante-mortem post-overdose results, given that vitreous humour magnesium probably reflects plasma magnesium
3. although the temporal relationship between magnesium administration and the fatal event might suggest a causal relationship, there was no toxicological or cardiological evidence of fatal magnesium poisoning.

Verdict: death by natural causes.

magnesium overdose may include drowsiness, unconsciousness, loss of muscle tone and respiratory and cardiac arrest. Plasma magnesium in healthy adults ranges from 0.7 to 1.3 mmol/L. Features of toxicity are reported at plasma magnesium concentrations of 3.5–5.0 mmol/L, and cardiorespiratory arrest may occur at concentrations of >8.5 mmol/L. During life, magnesium is distributed within cells (approximate cell:plasma distribution ratio 4:1), hence whole blood concentrations are somewhat higher than plasma concentrations. Doses of magnesium sulphate of up to 4 g (16 mmol magnesium ion) are sometimes given for control of convulsions.

Sodium

Hypernatraemia from sodium chloride (common salt) poisoning must be distinguished from hypernatraemia owing to dehydration from fluid loss or, very rarely, low fluid intake. This is especially important because sodium chloride poisoning may be the result of deliberate administration of large amounts of salt by a third party – salt is hardly ever encountered in self-poisoning episodes nowadays, although there were many deaths, especially in children, when saline emetics were used in an attempt to induce vomiting.

When a patient presents with hypernatraemia, diabetes insipidus and renal disease have to be excluded before salt poisoning can be diagnosed. Hypernatraemia caused by disease processes may be associated with obvious polyuria and polydipsia. Negative water balance can occur easily. This reduces solute excretion and induces hypernatraemic dehydration. Finally, there are rare individuals who develop hypodipsic hypernatraemia (see p. 44). A water deprivation test may be required if this disorder is suspected in life.

Common criteria used to diagnose salt (sodium chloride) poisoning focus on hypernatraemia, with high urinary concentrations of sodium and chloride. However, high urinary concentrations of sodium alone cannot distinguish salt poisoning from dehydration. The medical and legal implications of the two conditions are fundamentally different, so reliable ways to distinguish between them are needed. Be this as it may, both salt poisoning and dehydration (caused by neglect) may lead to criminal or civil action. Fractional excretions (the proportions of sodium and water filtered at the glomerulus that subsequently reach the urine), calculated from the sodium and creatinine concentrations of paired plasma and associated ('spot') urine samples, can distinguish the two situations (see Box 44.9). The values should be >2% in an individual who has been poisoned with salt and is volume replete and <1% when dehydrated but with viable renal tubules. Note that the units of measurement used must be the same for plasma and urine (plasma creatinine is usually reported in $\mu\text{mol/L}$, while urine creatinine is often reported in mmol/L).

POST-MORTEM BIOCHEMISTRY

Gradients that are maintained by active processes during life, such as that between intra- and extracellular potassium begin to break down soon after the occurrence of

BOX 44.9 Derivation of the fractional excretion of sodium (FE_{Na}) and water ($\text{FE}_{\text{H}_2\text{O}}$)

Fractional excretion of water, $\text{FE}_{\text{H}_2\text{O}}$

- The $\text{FE}_{\text{H}_2\text{O}}$ is the volume of electrolyte-free water that appears as urine compared with the amount filtered
- Thus, $\text{FE}_{\text{H}_2\text{O}} = V / \text{GFR}$
- Since $\text{GFR} = U_{\text{Cr}} \times V / P_{\text{Cr}}$
- $\text{FE}_{\text{H}_2\text{O}} = V \times P_{\text{Cr}} / U_{\text{Cr}} \times V$
- Simplifying, $\text{FE}_{\text{H}_2\text{O}} = P_{\text{Cr}} / U_{\text{Cr}}$ (then multiply by 100 to express as a percentage)

Fractional excretion of sodium, FE_{Na}

- The FE_{Na} is the amount of sodium lost in the urine compared to the amount filtered
- Thus, $\text{FE}_{\text{Na}} = \text{urinary sodium excretion} / \text{filtered sodium}$
- The urinary sodium excretion = $U_{\text{Na}} \times V$
- And the filtered sodium = $P_{\text{Na}} \times \text{GFR}$
- Since $\text{GFR} = U_{\text{Cr}} \times V / P_{\text{Cr}}$
- The filtered sodium = $P_{\text{Na}} \times U_{\text{Cr}} \times V / P_{\text{Cr}}$
- Therefore, $\text{FE}_{\text{Na}} = U_{\text{Na}} \times V \times P_{\text{Cr}} / P_{\text{Na}} \times U_{\text{Cr}} \times V$
- Simplifying, $\text{FE}_{\text{Na}} = U_{\text{Na}} / P_{\text{Na}} \times P_{\text{Cr}} / U_{\text{Cr}}$ (then multiply by 100 to express as a percentage)

P, plasma concentration; U, urine concentration; Cr, creatinine; GFR, glomerular filtration rate; V, urine flow rate.

Adapted from: Coulthard MG, Haycock GB

Distinguishing between salt poisoning and hypernatraemic dehydration in children. *BMJ* 2003; 326: 157–160.

BOX 44.10 Diabetic or alcoholic ketoacidosis?

A 56-year-old female had a history of chronic alcohol abuse and diabetes mellitus requiring insulin, probably secondary to an episode of alcohol-induced pancreatitis. Her husband found her collapsed one evening. This was a regular occurrence. He measured her capillary glucose concentration, and found it to be 1.5 mmol/L. He gave her an injection of 1 mg glucagon. She woke up, and so he went to bed. He awoke the following morning to find her dead in bed.

A blood sample was obtained post-mortem. The only notable finding was a raised β -hydroxybutyrate concentration (11.36 mmol/L). No ethanol was detected. It was not possible to measure glucose because of the condition of the sample.

At the inquest, it was thought that an individual who had abused alcohol chronically was unlikely to have a significant hepatic glycogen store, hence it was felt unlikely that a standard dose of glucagon would have led to excessive release of glucose. Thus, it was thought more likely that the ketoacidosis was due to the consequences of chronic alcohol abuse rather than to diabetic ketoacidosis.

hypoxic or anoxic damage, hence the possibility of both terminal and post-mortem change has to be evaluated when interpreting results. Most deaths that become the subject of post-mortem investigation occur outside hospital and it may be some days before the body is found and samples collected for analysis. Therefore, blood samples are invariably haemolysed to a greater or lesser degree and the likelihood of other changes, such as loss of labile analytes (e.g. glucose and insulin), is high (see Box 44.10). It should be remembered that most clinical

reference values are established for plasma or serum, and not haemolysed whole blood.

Vitreous humour

Vitreous humour is preferred to blood for most post-mortem biochemistry (Table 44.2), since it is thought to be far less susceptible to autolytic change, is less likely to be subject to post-mortem contamination by diffusion of drugs or other poisons that may be present at high concentration in the thorax or abdomen at death, and lies within the relatively protected environment of the eye socket. After death, however, potassium quickly leaks from the retina and hence vitreous potassium is not a reliable indicator of ante-mortem plasma potassium and is of minimal value in the diagnosis of exogenous potassium administration. The possibility of concurrent vitreous disease confounding the results must also be remembered.

When collecting vitreous humour, ideally both eyes should be sampled independently and the results reported separately. Potassium concentrations may differ by up to 2.3 mmol/L between the two eyes (in samples from non-putrefied bodies). Specimen contamination with retinal cells is a recognized source of falsely raised vitreous potassium concentrations. Hence, aspiration must be gentle

to avoid contamination with retinal fragments as far as possible. Differences may also be due to inappropriate sample handling. Vitreous humour is viscous, hence may require treatment such as centrifugation, heating, dilution or addition of hyaluronidase, to facilitate accurate pipetting. A vitreous potassium concentration >15 mmol/L suggests gross post-mortem decomposition.

Vitreous sodium and chloride concentrations may fall after death, at rates of up to 1 mmol/L per hour, whereas potassium increases at a rate of 0.14–0.19 mmol/L per hour. If the potassium concentration is <15 mmol/L, then the sodium and chloride concentrations are thought likely to reflect the situation at death. Urea and creatinine are relatively stable in post-mortem specimens. If vitreous sodium, chloride and urea are >155, >115 and >10 mmol/L, respectively, this may indicate ante-mortem dehydration. If the urea concentration is >20 mmol/L and creatinine >200 µmol/L with sodium and chloride concentrations within the normally accepted range, this indicates that uraemia was present before death.

A number of biochemical tests have been advocated for particular purposes, but remain little used (see Table 44.3). Attempts to employ the rate of rise of vitreous humour potassium concentration, in order to calculate the time of death, have largely been abandoned because of the

TABLE 44.2 Post-mortem biochemistry interpretation

Analyte	Matrix	Acceptable post-mortem range	Interpretation of raised concentration
Glucose	Vitreous humour	After death, vitreous humour glucose falls rapidly, therefore any detectable glucose requires investigation	(Drug induced) hyperglycaemia, diabetic ketoacidosis, stress response (interpret in conjunction with lactate)
β-Hydroxy-butyrate	Blood, vitreous humour	0.1–1.0 mmol/L	Fasting, prolonged alcohol abuse, diabetic ketoacidosis, stress response (e.g. hypothermia)
HbA _{1c}	Blood	27–67 mmol/mol	Poor long-term (2–8 weeks) blood glucose control
Creatinine	Vitreous humour	<100 µmol/L	Poor renal function; high protein intake; large muscle mass; heat shock
Lactate	Vitreous humour	<10 mmol/L	Interpret in conjunction with glucose
Sodium	Vitreous humour	135–145 mmol/L	Salt poisoning; dehydration (interpret in conjunction with creatinine and urea)
Chloride	Vitreous humour	95–105 mmol/L	Salt poisoning; dehydration (interpret in conjunction with creatinine and urea)
Urea	Vitreous humour	<10 mmol/L	Poor renal function; upper GI haemorrhage
Tryptase	Blood	<100 µmol/L	Anaphylactic shock

TABLE 44.3 Post-mortem biochemistry: some little used/unvalidated tests

Analyte	Matrix	Suggested role in diagnosis of:
Adrenaline:noradrenaline ratio	Urine	Hypothermia
Chymase	Blood	Anaphylactic shock
Chromogranin A	Serum, cerebrospinal fluid	Hypothermia
C-reactive protein (CRP)	EDTA blood, liver	Recent infection, trauma, burns, ketoacidosis, necrosis. Diagnosis of sepsis if measured soon after death
Fructosamine	Vitreous humour	Diabetic ketoacidosis
Hypoxanthine	Vitreous humour	Time of death
Lactate	Vitreous humour	If very high, may indicate lactic acidaemia (but may be formed perimortem)
Thyroglobulin, free tri-iodothyronine (fT ₃)	Blood	Neck trauma (e.g. strangulation)
Troponin	Pericardial fluid	Myocardial damage present prior to microscopic changes

inherent uncertainty of this method, even when vitreous humour hypoxanthine is also analysed, with the aim of increasing the accuracy of the procedure. Use of synovial fluid, as an alternative to vitreous humour, for potassium measurement for time-of-death estimation, has also been advocated. It has been claimed that regression analysis of vitreous humour hypoxanthine, potassium and urea concentrations can give a reliable time-of-death estimate, taking into account the cause of death. There are reports of immunochemical detection of glucagon in pancreatic cells and of calcitonin in thyroid C-cells being used to indicate time of death, but these methods are experimental at best. Serial monitoring of C-reactive protein (CRP), procalcitonin, interleukin-6, interleukin-1 β , soluble interleukin-2 receptor or lipopolysaccharide binding protein in the hours after death, has been suggested as an aid to the diagnosis of sepsis, but has not been widely adopted.

The measurement of markers of heart muscle damage (e.g. myoglobin, creatine kinase, troponin) as an adjunct to searching for evidence of microscopic changes in samples of myocardium post-mortem, has been investigated with the aim of improving the diagnosis of ante-mortem ischaemic damage. However, the only post-mortem fluid with which there has been some success in this approach, has been pericardial fluid; concurrent post-mortem changes limiting the value of fluids such as blood. Unfortunately, as microscopy has always been viewed as the definitive method for the diagnosis of ante-mortem heart damage, the use of biochemical markers in cases where death has occurred prior to microscopic changes becoming apparent has not been fully investigated. Hence, the interpretation of results remains problematic.

SPECIFIC DIAGNOSTIC PROBLEMS

Anaphylaxis/anaphylactoid reactions

Many drugs may precipitate anaphylaxis or anaphylactoid reactions. Mast-cell tryptase is an indicator of anaphylactic shock. Measurement of blood chymase has also been suggested in this context. Although a raised serum or plasma tryptase activity can supply important supporting evidence in the diagnosis of anaphylaxis, it cannot be used as the sole criterion for the post-mortem diagnosis of anaphylaxis as a cause of death, since there is overlap with results from people who die from causes other than anaphylaxis.

Diabetes

Blood and vitreous humour glucose concentrations generally fall rapidly after death and are thus an unreliable guide to ante-mortem glucose concentration. However, since glucose undergoes anaerobic glycolysis to lactate, some investigators have suggested that measuring the sum of vitreous humour glucose and lactate may give a better estimation of the glucose concentration at the time of death. A potentially complicating factor is that lactate may increase perimortem. Moreover, some drugs that are themselves unstable in biological systems, for example

ethanol and insulin, may cause fatal hypoglycaemia, hence compounding the difficulties that may be encountered in establishing a cause of death.

An elevated post-mortem blood HbA_{1c} concentration might indicate poor glucose control during life. This suggestion may be supported by an elevated vitreous humour glucose concentration. In patients with type 1 diabetes, the presence of ketones (principally acetone and its metabolite 2-propanol, but also acetoacetate) may indicate the presence of diabetic ketosis prior to death, especially if considered together with the blood β -hydroxybutyrate (BHB) concentration. Schemes to aid interpretation of vitreous glucose concentrations, similar to that in [Figure 44.1](#), are often cited, but even these may not give clear guidance as to the likely situation perimortem in some cases.

Measurement of ketones in blood or vitreous humour is recommended in unexplained deaths in chronic alcoholic as well as diabetic patients. Prolonged poor nutrition and use of alcohol (common in chronic alcoholics) promote the accumulation of ketones (acetone, butanone) and BHB, and often, elevated ketone concentrations (acetone >20 mg/L) are the only notable post-mortem finding. A practical scheme for investigating such deaths is given in [Figure 44.2](#).

Drowning

Haemodilution is likely in victims of fresh-water drowning, but not in those drowned in salt water. Thus, concentrations of ethanol, drugs or other analytes measured in the blood of victims of fresh-water drowning may be misleadingly low. Conversely, haemoconcentration is likely if a cadaver has been dehydrated, for example by heat or by mummification. In all such cases, measurement of blood haemoglobin may give an estimate of the magnitude of haemodilution/concentration, if it has not been degraded by heat or by prolonged storage.

It has been suggested that haemodilution in fresh-water drowning produces a lower chloride concentration in left heart blood as compared with right heart blood, while in salt water drowning, haemoconcentration and chloride ion absorption is said to produce the opposite result. High concentrations of magnesium in left heart blood compared with right heart blood may reflect magnesium absorption from salt water. Trace element analysis has been proposed as an aid to discriminate fresh-water drowning from salt-water drowning, but seems to offer poor discriminatory power and other (non-biochemical) investigations, notably looking for species of diatoms, are more helpful in such cases.

Hypothermia/hyperthermia

Cold may be a significant factor in a death, but there are no specific biochemical markers that can be used to confirm a diagnosis of fatal hypothermia. Indications of ante-mortem cold stress include high urinary adrenaline and noradrenaline concentrations and Wischnewsky spots (small gastric mucosal bleeds). An increased adrenaline:noradrenaline ratio may be a better indicator of ante-mortem hypothermia than the separate

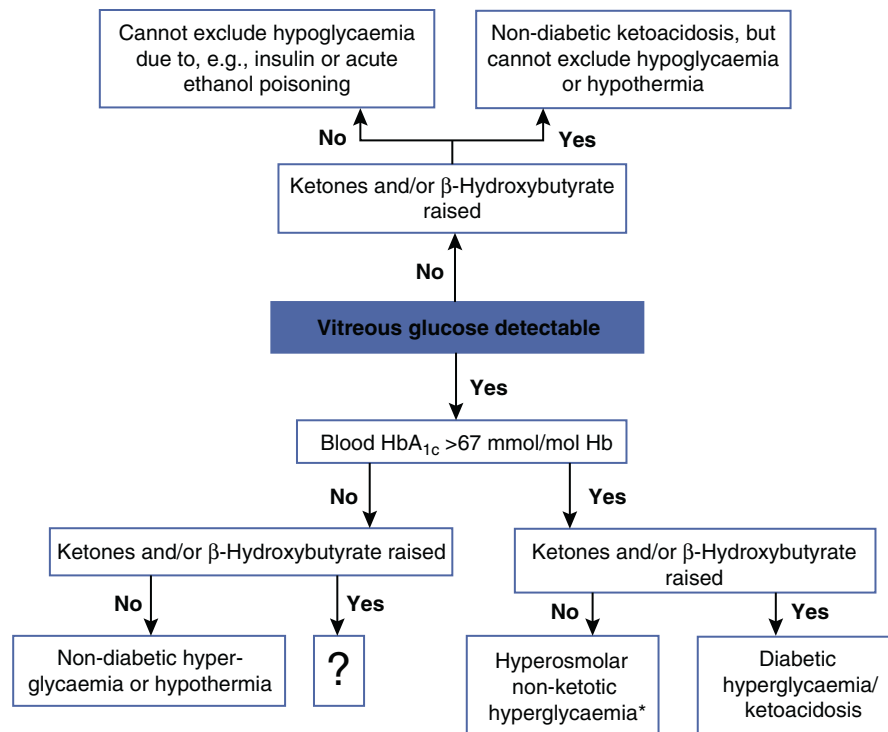


FIGURE 44.1 ■ Aid to the interpretation of post-mortem vitreous humour glucose and blood β -hydroxybutyrate concentrations. *Requires confirmation from vitreous humour sodium and urea/creatinine.

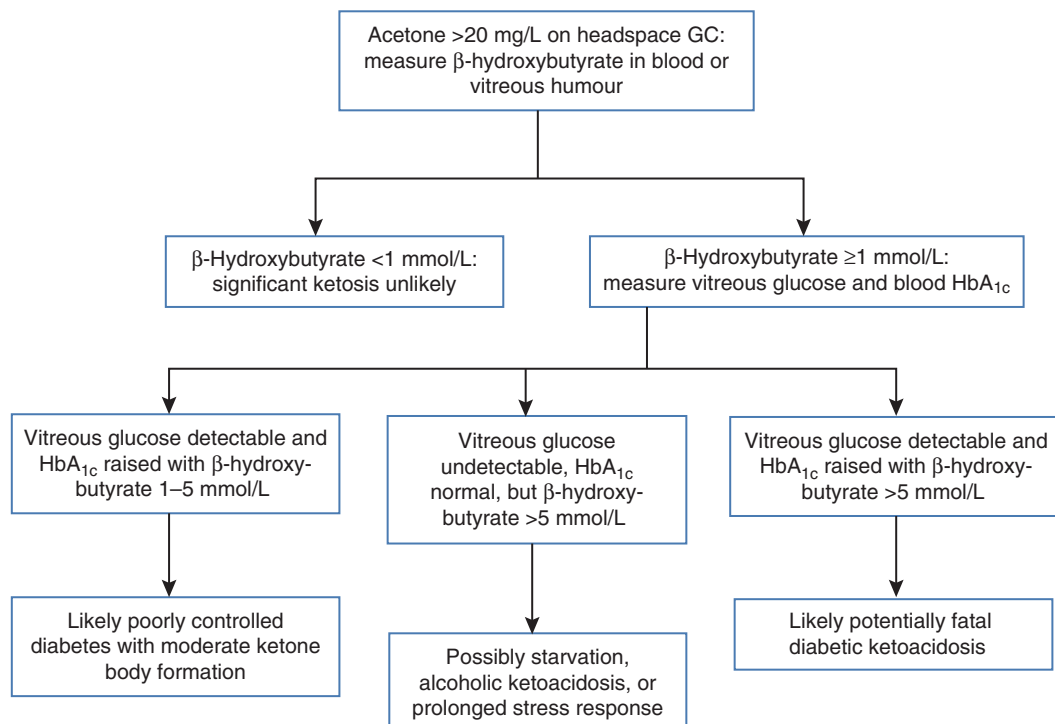


FIGURE 44.2 ■ Suggested scheme for investigation of unexpected deaths in patients in whom diabetes or alcoholism is a possibility.

catecholamine results. Ketone and glucose concentrations may also be elevated, and in persistent hypothermia, there may be electrolyte disturbances and metabolic acidosis. A role for measurement of chromogranin A has been postulated.

Fatal hyperthermia (heat stroke) often involves multiple organ dysfunction, including skeletal muscle damage without marked inflammatory responses. It has been suggested that isolated elevation of serum creatinine may help in diagnosis.

Inflammation

Blood CRP concentration peaks within 6 h of a stimulus and CRP is stable in post-mortem samples. Liver is said to be a good alternative specimen if blood is not available. Interpretation of results can be difficult, however, as there are many causes of a raised blood CRP in addition to inflammation (see Table 44.3).

Sudden death

Sudden unexpected death in infancy (SUDI, also known as sudden infant death syndrome (SIDS), Table 44.4), may be an acute manifestation of an inborn error of metabolism, particularly a fat oxidation defect, and therefore these should be excluded. Sudden arrhythmic death syndrome (also known as sudden adult death syndrome (SADS)) is thought likely often to have a cardiac origin

TABLE 44.4 Some biological samples required when investigating sudden unexpected death in infancy (SUDI)

Sample (volume)	Handling	Test
Blood (serum) (1–2 mL)	Centrifuge, store serum at –20°C	Toxicology
Urine (20 mL if possible)	Store at –20°C	Toxicology and specialized tests for inherited metabolic diseases
Blood from Guthrie card	Ensure circle is filled. Do not put in plastic bag	Tests for inherited metabolic diseases
Skin biopsy	After discussion with paediatrician and laboratory	Tests for inherited metabolic disease, e.g. fibroblast enzyme activity
Muscle biopsy	After discussion with paediatrician and laboratory	If history suggests mitochondrial disorder

Further details: Royal College of Pathologists and Royal College of Paediatrics and Child Health (2004).

such as a fatal arrhythmia, but a full toxicological analysis is required to exclude recent use of not only illicit drugs (such as cocaine and other stimulants), but also therapeutic drugs (e.g. tricyclic antidepressants and antipsychotics) that may increase the risk of a fatal cardiac arrhythmia.

Further reading

Up-to-date reviews on post-mortem biochemistry

- Madea B. Sudden death, especially in infancy – improvement of diagnoses by biochemistry, immunohistochemistry and molecular pathology. *Legal Medicine (Tokyo)* 2009;11(Suppl. 1):S36–42.
- Maeda H, Ishikawa T, Michiue T. Forensic biochemistry for functional investigation of death: concept and practical application. *Legal Medicine (Tokyo)* 2011;13:55–67.
- Palmiere C, Lesta Mdel M, Sabatasso S et al. Usefulness of post-mortem biochemistry in forensic pathology: illustrative case reports. *Legal Medicine (Tokyo)* 2012a;14:27–35.

Practical guidance on post-mortem samples and sampling

- Dinis-Oliveira RJ, Carvalho F, Duarte JA et al. Collection of biological samples in forensic toxicology. *Toxicol Mech Methods* 2010;20:363–414.
- Flanagan RJ, Connally G, Evans JM. Analytical toxicology: guidelines for sample collection post-mortem. *Toxicol Rev* 2005;24:63–71.
- Royal College of Pathologists and Royal College of Paediatrics and Child Health. Sudden unexpected death in infancy. A multiagency protocol for care and investigation. London: RCP and RCPCH; 2004. <http://www.rcpath.org/NR/rdonlyres/30213EB6-451B-4830-A7FD-4EEFF0420260/0/SUDIreportforweb.pdf> [Accessed 30.09.13].

Clinical and post-mortem diagnosis of disorders of glucose metabolism

- Bouलगnon C, Garnotel R, Fornes P et al. Post-mortem biochemistry of vitreous humor and glucose metabolism: an update. *Clin Chem Lab Med* 2011;49:1265–70.
- Elliott S, Smith C, Cassidy D. The post-mortem relationship between beta-hydroxybutyrate (BHB), acetone and ethanol in ketoacidosis. *Forensic Sci Int* 2010;198:53–7.
- Hess C, Musshoff F, Madea B. Disorders of glucose metabolism: post-mortem analyses in forensic cases: part I. *Int J Legal Med* 2011;125:163–70.
- Hockenhull J, Dhillon W, Andrews R et al. Investigation of markers to indicate and distinguish death due to alcoholic ketoacidosis, diabetic ketoacidosis and hyperosmolar hyperglycemic state using post-mortem samples. *Forensic Sci Int* 2012;214:142–7.
- McGuire LC, Cruickshank AM, Munro PT. Alcoholic ketoacidosis. *Emerg Med J* 2006;23:417–20.
- Musshoff F, Hess C, Madea B. Disorders of glucose metabolism: post-mortem analyses in forensic cases. Part II. *Int J Legal Med* 2011;125:171–80.
- Palmiere C, Sporkert F, Werner D et al. Blood, urine and vitreous isopropyl alcohol as biochemical markers in forensic investigations. *Leg Med (Tokyo)* 2012b;14:17–20.

Reviews on post-mortem diagnosis of anaphylaxis, hyperthermia/hypothermia and sepsis

- Da Broi U, Moreschi C. Post-mortem diagnosis of anaphylaxis: a difficult task in forensic medicine. *Forensic Sci Int* 2011;204:1–5.
- Tsokos M. Post-mortem diagnosis of sepsis. *Forensic Sci Int* 2007;165:155–64.
- Yoshida C, Ishikawa T, Michiue T et al. Post-mortem biochemistry and immunohistochemistry of chromogranin A as a stress marker with special regard to fatal hypothermia and hyperthermia. *Int J Legal Med* 2011;125:11–20.

Note

Page numbers suffixed by 'b', 'f' and 't' refer to boxes, figures and tables respectively.

A

- Abdomen, acute 227–230
causes 228, 228b
- Abdominal pain, in acute porphyria 538, 541
- Abetalipoproteinemia 689, 700, 722t, 724
- ABO blood group 510, 510t
laboratory tests 511, 511t
- Abortion, spontaneous 443
- Acanthosis nigricans 294–295
- Acarbose 317
- Accidental poisoning 788. *See also* Poisoning
- Accuracy 10–11
- Acetaldehyde 257
- Acetaminophen. *See* Paracetamol
- Acetoacetate, in ketoacidosis 77, 299, 327
- Acetyl-CoA, metabolism 647–649, 650f, 657
- Acetylcysteine 794t, 796
- N-Acetyl β -D-glucosaminidase (NAG) 161–162, 162b
- N-Acetylglutamate synthetase (NAGS) deficiency 463f, 478
- N-Acetyl-p-benzoquinoneimine (NAPQI) 795, 795f
- Acetylsalicylic acid. *See* Aspirin
- Aches
in articular disorders 638–639. *See also* Pain
- Acholuric jaundice 245
- Acidaemia 65–66
methylmalonic 479, 480f
propionic 479, 480f
- Acid–base disturbances. *See* Hydrogen ion homeostasis, disorders
- Acid–base status
in acute kidney injury 140
assessment 73–74
in chronic kidney disease 143–144
data interpretation 85–86, 85f
potassium distribution and 32. *See also* Hydrogen ion homeostasis
- Acid-labile subunit (ALS), in tumour-related hypoglycaemia 282, 343
- Acidosis
anion gap 74
definition 65–66
dilutional 78
effect on TmP/GFR 112
lactic. *See* Lactic acidosis
metabolic. *See* Acidosis, non-respiratory
osteomalacia and 623
potassium secretion and 33
in renal disease 78–80
acute kidney injury 78–79, 140
chronic kidney disease 78–79, 143
renal hypokalaemic 54–55, 54b, 57
renal tubular. *See* Renal tubular acidosis (RTA)
respiratory. *See* Acidosis, respiratory
systemic effects 75–76
urinary ammonium excretion 71
- Acidosis, non-respiratory 74–80
biochemical characteristics 75
buffering 74
causes 74, 75b, 76–80
compensatory responses 73, 74–75
in diabetic subjects 330. *See also* Diabetic ketoacidosis (DKA)
differential diagnosis 469b
hyperventilation 73, 74–75
in inherited metabolic disease 469, 469b, 473–474
interpretation of acid–base data 85, 85f
management 76
in neonates 488
osteomalacia and 623
poisoning-related 790t
renal hydrogen ion excretion 75. *See also* specific causes
- Acidosis, respiratory 80–82
biochemical characteristics 81
buffering 80–81
causes 81b
compensatory responses 80–81
hyperventilation 81
interpretation of acid–base data 85, 85f
management 81–82
renal hydrogen ion excretion 81
systemic effects 81
- Acid phosphatase, tartrate-resistant 613
- Acidurias, amino 136, 170–172
- Acid α -glucosidase 654–655
- Acinus 232–233, 233f
- Acquired immunity 560. *See also* Adaptive immune system
- Acromegaly 362–363
diagnosis 362
management 362–363
monitoring response to therapy 363
presentation 362–363
hypertension 748–749
peripheral neuropathy 693
secondary diabetes 294
- ACTH. *See* Adrenocorticotrophic hormone (ACTH)
- Actin 646, 647f, 740
- Action limits 16
- Activated charcoal 794, 797, 798, 803
- Activated partial thromboplastin time (APTT) 508, 509t
- Active transport, in renal tubule 169f
- Activin 436–437
- Acute abdomen. *See* Abdomen, acute
- Acute chest syndrome 556
- Acute confusional state. *See* Delirium
- Acute coronary syndrome (ACS) 738, 739f, 744
non-ST elevation 739f, 744
treatment 744, 744f. *See also* Myocardial damage, acute; Myocardial infarction
- Acute dilutional hyponatraemia 46–47, 47f, 50
causes 47b
laboratory investigation 50
management 51. *See also* Hyponatraemia
- Acute fatty liver of pregnancy (AFLP) 263, 445, 465–466
- Acute intermittent porphyria (AIP) 534t, 537t, 538
hepatocellular carcinoma risk 539
homozygous 542
molecular genetics 534–537, 537t
pathophysiology 538
screening 541t. *See also* Porphyria(s), acute
- Acute kidney injury (AKI) 136–141
acid–base balance 140
causes 136–137, 137b
chronic kidney disease vs. 136
in cirrhosis 260–261
differential diagnosis 261, 261t
classification 136–137
clinical features 136t
fluid and electrolyte balance 140
hepatorenal syndrome. *See* Hepatorenal syndrome (HRS)
hypertension in 759
intrinsic 137b, 138–139
diagnosis 138, 138t
management 140–141
general 140
renal replacement treatment 141
metabolic consequences 140–141
acidosis 78–79, 140
hypercalcaemia 103t, 140
hyperkalaemia 140
metabolic features 136t
nutrition and 140
poisoning-related 793
postrenal (obstructive) 137b, 139
prerenal 137–138, 137b
diagnosis 137–138, 138t
management 138
in setting of chronic kidney disease 139, 139b
- Acute liver failure (ALF) 253
amino acid metabolism 235
assessment of prognosis 249
carbohydrate metabolism 234
laboratory features 253
liver transplantation 253
monitoring 246
in pregnancy 263
progression from viral hepatitis 252–253
- Acute myeloid leukaemia (AML) 505f
- Acute myelomonocytic leukaemia, renal potassium loss 56
- Acute phase proteins 408, 571–572
functions 572t
properties 572, 572t
psychiatric disorder investigation 676t
- Acute phase reactants, liver function tests 243
- Acute phase response, investigation/markers 598, 644

- Acute phosphate deficiency syndrome 116, 116t
 Acute rejection 599b
 Acute renal failure. *See* Acute kidney injury
 Acute sodium loading 38
 Acute tubular necrosis (ATN) 138–139
 natural history 139
 pathogenesis 138, 139f
 Acylcarnitine profiling 473
 Acyl-CoA:cholesterol acyltransferase (ACAT)
 718
 Acyl-CoA:cholesterol acyltransferase 2
 (ACAT2) 711, 718
 Acyl-CoA dehydrogenases 647–649, 649f
 Acylglycines, measurement in urine 657
 Adaptive immune system 562–570, 562t
 antigen recognition 566–570
 cells 562, 562t, 563t, 564f. *See also* White
 blood cells (WBCs)
 cellular immune activation 570
 lymphoid tissue 562, 562t, 563f
 soluble mediators 562, 562t
 Addison disease 59, 583
 clinical features 369–370
 depression 679
 hypercalcaemia 103t
 hypoglycaemia 345
 salt wasting 35
 Addisonian crisis 369–370
 hypoglycaemia in 345
 Adenohypophysis. *See* Pituitary gland,
 anterior lobe
 Adenomas
 pituitary. *See* Pituitary adenomas
 toxic. *See* Toxic adenoma
 Adenomatous primary hyperparathyroidism
 101, 102t, 103–104
 Adenosine deaminase (ADA) deficiency 481,
 528
 Adenosine phosphoribosyl transferase
 (APRT) deficiency 178
 Adenosine triphosphate (ATP) 109–110
 generation in red blood cells 517, 518f
 defective 528
 muscle contraction role 646
 cardiac muscle 740–741
 Adenoviruses, gene therapy vectors 870
 Adenylate kinase 647
 Adherence, drug 767–768
 therapeutic drug monitoring 772
 Adipocyte lipid-binding protein (ALBP) 717
 Adipose tissue
 insulin actions 280t
 lipolysis 713
 Adipose triglyceride lipase (ATGL) 713
 Adipsic hypernatraemia 44
 Adjustable gastric band (AGB) 206, 206f
 Adolescents, with disorder of sex
 development 419–420, 419f
 Adrenal glands
 anatomy 352
 disease 369–370
 autoimmune 583
 hormone replacement therapy,
 monitoring 370–371. *See also specific
 diseases*
 function, depression and 678–679, 680
 function tests
 steroid therapy and 358–359. *See also*
 Adrenocorticotrophic hormone
 (ACTH)–adrenal axis, assessment
 physiology 352, 353f
 Adrenal incidentaloma 762, 764
 assessment 370
 Adrenaline
 secretion, hypoglycaemia response 334
 stress response 405, 407–408
 Adrenaline:noradrenaline ratio, post-mortem
 biochemistry 879t, 880–882
 Adrenal insufficiency, primary. *See* Addison
 disease
 Adrenal medulla, stress response 405
 Adrenal suppression, steroid therapy-
 associated, assessment 358–359
 Adrenarche 414
 premature 429
 Adrenergic receptors 407
 Adrenocorticotrophic hormone (ACTH)
 350–351
 deficiency 352, 354, 358, 360
 hypoglycaemia due to 345
 investigation 353–354, 357, 358
 isolated 369
 in depression 678–679
 petrosal sinus sampling 365–366
 plasma measurements, Cushing syndrome
 364, 365
 in post-traumatic stress disorder 679
 secretion 350–351
 ectopic 364, 761t, 814, 815t. *See also*
 Adrenocorticotrophic hormone
 (ACTH)–dependent Cushing
 syndrome
 stress response 9, 404
 test (short tetracosactide test) 354, 355,
 358, 372
 Adrenocorticotrophic hormone (ACTH)–
 adrenal axis, assessment
 clinical approach 358–359
 dynamic tests of function 354–355
 reproducibility 355
 Adrenocorticotrophic hormone
 (ACTH)–dependent Cushing
 syndrome
 clinical context 364–365
 differential diagnosis 364, 364f. *See also*
 Cushing syndrome
 Adrenoleukodystrophy (ALD) 465
 Adrenoleukodystrophy protein (ALDP) 465
 Adrenomyeloneuropathy (AMN) 465, 689
 Advanced glycation end products (AGEs)
 296–297
 Adynamic bone disease 625
 Aerobic exercise tests 655
 Affective disorders 675, 675t. *See also specific
 affective disorders*
 Affinity maturation 561
 AFP (α -fetoprotein)
 Down syndrome screening 443
 liver function tests 242
 plasma concentrations
 in children 837–838
 effect of gestational age on 443, 443f
 as tumour marker 264, 822t, 826t
 gastric cancer 831
 germ cell tumours 832, 833–834, 833f,
 834f
 hepatocellular carcinoma 264, 835–836,
 838
 Age
 body fluid distribution in relation to 28t
 cardiovascular disease risk and 751, 751f
 effect on test results 8, 8t
 effect on urinary protein excretion 155
 Age-related changes
 alkaline phosphatase 8, 493, 610f, 632
 bone 609, 615
 cerebrospinal fluid proteins 663, 663t
 immunoglobulins 575–576, 576f
 osteocalcin 610f, 632
 plasma phosphate concentration
 113, 113t
 thyroid function 379
 Agouti-related peptide (AgRP) 203
 Alanine aminotransferase (ALT)
 in acute hepatitis 251
 in alcoholic hepatitis 257
 analytical goals 11t
 analytical variation 11t
 biological variation 11t
 in cerebrospinal fluid 667
 critical difference 15t
 liver function tests 240
 in muscle disease 653. *See also*
 Aminotransferases
 Alarmins 406
 Albright hereditary osteodystrophy 491
 Albumin 2, 242
 analytical goals 11t
 analytical variation 11t
 biological variation 11t
 cerebrospinal fluid 663, 663f, 663t
 in assessment of blood–brain barrier
 permeability 664
 clearance 154t
 critical difference 15t
 glomerular filtration rate 153–154, 160
 liver function tests 242
 oncotic pressure 28
 plasma concentration
 in chronic hepatitis 254
 in malnutrition 194–195
 in pregnancy 446
 in protein–energy malnutrition 201
 reabsorption 154
 sodium excess and 37
 thyroid hormone transport 376
 urinary excretion 152, 155, 155t
 age-related differences 155
 in diabetes 322
 diet and 155
 glomerular *vs.* tubular disease 160, 160t
 in pregnancy 156. *See also*
 Microalbuminuria
 urine, measurement 165–166
 investigation of stone formers 178
 Alcohol
 acid–base disturbances 77
 cardiovascular disease risk and 756
 effects 800
 on biochemical variables 8t
 on bone mineral density 615
 on protein metabolism 258
 hypertriglyceridaemia association 731
 metabolism 257, 801f
 metabolites, measurement 258
 poisoning 685, 792t, 799–800, 801f
 withdrawal syndrome 800. *See also* Alcohol
 misuse
 Alcoholic hepatitis 251, 254t, 257, 800
 Alcoholic ketoacidosis 76–77, 330, 800
 Alcoholic liver disease 257–258, 800
 biochemical abnormalities 257–258
 γ -glutamyltransferase activity 241, 257
 iron overload 257
 differential diagnosis 254t, 255
 laboratory tests 258
 liver pathology 257–258
 Alcoholic steatosis 257–258
 Alcohol-induced hypoglycaemia 321, 346–347
 Alcohol misuse
 acute pancreatitis and 228, 229t
 diabetes mellitus and 293
 thiamin deficiency and 686. *See also*
 Alcohol
 Aldolase B deficiency 464
 Aldose reductase 296
 diabetic neuropathy pathophysiology
 691–692

- Aldosterone 28–29, 29f
 effect of posture 9
 effect of potassium 33
 menstrual cycle and 28
 in pre-eclampsia 38
 reference ranges 766
 in stress response 406
 in syndromes of hypoaldosteronism 59–60, 59t
- Aldosterone-producing adenoma (APA) 760–761
 localization 762. *See also* Hyperaldosteronism
- Aldosterone:renin ratio 762
 confounding factors 762
 protocol for investigation 765–766
- Aldosteronism. *See* Hyperaldosteronism
- Alendronate 617–618
- Alfacalcidol 98, 622, 623, 627
- Alkalaemia 65–66
- Alkaline phosphatase (ALP) 239, 609
 in acute hepatitis 251
 age-related changes 8, 493, 610f, 632
 analytical goals 11t
 analytical variation 11t
 biological variation 11t
 bone 607
 as turnover marker 609–610, 610f, 626, 630
 in chronic hepatitis 254
 in chronic kidney disease–mineral and bone disorder 626
 critical difference 15t
 in hypophosphatasia 624
 in infants 492–493
 iso-enzymes 239
 analysis 240, 240f
 liver function tests 239–240, 245, 247
 enhancement of specificity 240, 240f
 in osteopenia of prematurity 492, 492t
 in Paget disease 630
 parenteral nutrition-associated changes 264
 placental, as tumour marker 823t, 832
 in pregnancy 262–263, 446
 in primary biliary cirrhosis 256
 in primary sclerosing cholangitis 256–257
 sex-related changes 493
- Alkalosis
 effect on TmP/GFR 112
 metabolic. *See* Alkalosis, non-respiratory
 post-hypercapnic 81, 83
 potassium secretion and 33
 renal hypokalaemic 54b, 55–56, 57
 respiratory. *See* Alkalosis, respiratory
 systemic effects 83
- Alkalosis, non-respiratory 75
 biochemical characteristics 83
 buffering 82
 causes 82b, 83–84
 compensatory responses 82–83
 hypoventilation 82
 interpretation of acid–base data 85, 85f
 management 83
 potassium depletion 82–83
 renal bicarbonate excretion 82–83
 systemic effects 83
- Alkalosis, respiratory 84–85
 biochemical features 84
 causes 84b
 compensatory responses 84
 hypophosphataemia 116
 in inherited metabolic disease 469
 management 84–85
 systemic effects 84
- Alkapton 643
- Alkaptonuria 482, 482f
 articular manifestations 643
- Allan–Herndon–Dudley syndrome 377
- Alleles 848
 definition 871
- Allele-specific amplification 855–856, 856f
- Allele-specific oligonucleotide (ASO)
 technique 855, 856f
 reverse method 856–857
- Allergies 580–582
 clinical features 580
 investigation 580–581
 IgE measurement 581, 581t, 582t
 indications 581t
 patient history 580–581, 580t
 skin prick testing 581, 581t
 prevalence 580
 therapeutic diets 209t. *See also* Anaphylaxis
- Allopurinol 641
 Allopurinol loading test 475
- Alpha-glucosidase inhibitors 317
- Alpha-subunit, measurement 386
- Alpha thalassaemia 552–553, 552t, 558
- α -Blockers, hypertension management in
 diabetes mellitus 309
- Alprostadil 459
- Alstrom syndrome 202t
- Alternative pathway, complement activation 570, 571f
- Aluminium
 plasma concentrations, measurement 626–627, 635
 retention in chronic kidney disease–mineral and bone disorder 625
 clinical features 625
 investigation 626–627, 635
 toxicity, treatment 628
- Aluminium hydroxide 114–115
 toxicity 148–149
- Alveolar gas equation 87
- Alveolar ventilation 87
- Alzheimer disease 665, 670, 677
 apolipoprotein E and 710, 719
 cerebrospinal fluid analysis 670–671
 genetics 678
 markers 678
- Amadori products 296–297
- Amadori rearrangement 300–301
- Amanitins 806
- Amenorrhoea 439–440
 causes 615, 615b
 depression 680
 hypothalamic 353, 356–357, 368f, 369
 investigations, in adolescent girls 419, 419f
- Amphetamine abuse 792t, 801, 801t
- Amikacin, therapeutic drug monitoring 781
- Amino acid(s)
 essential 183, 183b
 metabolism. *See* Amino acid metabolism
 non-essential 183, 183b
 parenteral nutrition 211
 plasma analysis, in inherited metabolic diseases 472, 472b
 renal tubular handling 168, 170
 transporters, L-type 377
 urinary analysis, in inherited metabolic diseases 472, 472b
- Amino acid disorders 472, 472t
 investigation 472, 472b
 primary 472
 renal 472, 472t. *See also* specific disorders
- Amino acid metabolism 69–70
 disorders. *See* Amino acid disorders
 hepatic 235, 235f
- Aminoacidurias 136, 170–172
- Aminoglycoside antibiotics
 therapeutic drug monitoring 770–771, 773, 781
 toxicity 161, 781
- Aminoguanidine 692
- 5-Aminolaevulinic acid (ALA) dehydratase 533, 535f
- 5-Aminolaevulinic acid dehydratase deficiency porphyria (ADP) 534t, 542
 biochemical findings 537t, 542
 molecular genetics 537t
- 5-Aminolaevulinic acid synthase (ALAS) 533, 535f
- Aminotransferases, liver function tests 240–241, 245, 247–248
 in acute hepatitis 251
 in alcoholic hepatitis 257
 in chronic hepatitis 254
 in hepatitis B infection 255
 parenteral nutrition-associated changes 264
 in primary sclerosing cholangitis 256–257
- Amiodarone, therapeutic drug monitoring 777
- Amiodarone-induced thyroid dysfunction 390t, 393
- Amitriptyline, therapeutic drug monitoring 780
- Ammonia
 buffering 67
 intoxication, urea cycle defects 462, 469
 plasma measurements, in inherited metabolic diseases 469, 470
 production and clearance 235, 235f
- Ammonium
 buffering 67
 urinary excretion 67, 71, 72f
 non-respiratory acidosis 75
 respiratory acidosis 81
- Amniocentesis 444–445, 444f, 444t, 476, 860
- Amplification refractory mutation system (ARMS), genetic analysis in cystic fibrosis 863–864, 864f
- Amylase
 clearance 154t
 elevated activity, non-pancreatic causes 218t
 pancreatic function testing 217–218
 acute pancreatitis 228–229
 salivary 221
- Amylin 288
- Amylo-1:6-glucosidase 654–655
- Amyloid A, properties 572t
- Amyloidosis 596–597, 597t, 697–699
 biochemical abnormalities 696b
- Amyloid precursor protein (APP) 678
- Amyloid protein β 665–666
 deposition in Alzheimer disease 678
- Amylopectin 182f, 220–221
- Amylose 182f, 220–221
- Amyotrophy, diabetic 324
- Anaemia 517
 aplastic 524
 associated with increased red cell loss 524–529. *See also* Haemolytic anaemia
 associated with reduction in red cell production 517–524
 acquired causes 524
 inherited causes 524
 nutritional deficiencies 517–520
 vitamin B₁₂ and folate deficiency. *See* Megaloblastic anaemia
 of chronic disease 524, 643–644, 816, 816t
 in chronic kidney disease 143, 147, 149
 classification 517
 definition 501, 517

- Anaemia (*Continued*)
 in erythropoietic protoporphyria 547
 in gastric and small bowel disease 220
 haemolytic. *See* Haemolytic anaemia
 in hypothyroidism 395
 investigation in immunological diseases 601*t*
 iron deficiency. *See* Iron deficiency anaemia
 leukoerythroblastic 524
 in malignancy 816, 816*t*
 megaloblastic. *See* Megaloblastic anaemia
 morphological changes in red blood cells 501–503
 pernicious. *See* Pernicious anaemia
 in rheumatoid arthritis 643–644
 sickle cell. *See* Sickle cell disease
 symptoms 517
- Anaerobic glycolysis. *See* Glycolysis
- Anaesthesia, porphyria patients 542
- Analbuminaemia, congenital 37
- Analgesics
 narcotic, antidote 794*t*
 therapeutic drug monitoring 777
 tubular damage 161
- Analytical factors, affecting test results 10–12
- Analytical goals 11*t*, 12
- Analytical precision 11, 11*t*
- Analytical range 10–12
- Analytical variation 9–10, 11, 11*t*, 15
- Anaphylactic shock 574–575
- Anaphylactoid reactions 581–582
 post-mortem biochemistry 880
- Anaphylatoxins 157, 571*t*
- Anaphylaxis 581–582
 clinical features 582*t*
 investigation 582, 601*t*
 mediators 572*t*, 581–582
 post-mortem biochemistry 880
 treatment 581–582
- Androgen(s) 262
 deficiency/resistance, XY disorder of sex development 425, 426*t*
 effect on prostate cancer 820
 excess
 polycystic ovary syndrome 440
 XX disorder of sex development 423–425, 424*t*
 in females 439
 secretion 352
 sensitivity, assessment 428
- Androgen insensitivity syndromes 456
 complete 426*t*, 427–428, 456
 partial 419–420, 426*t*, 427–428, 456
- Androstenedione
 biosynthesis 437–438, 438*f*
 in females 439
 excess 440
 human chorionic gonadotrophin stimulation test 422
- Angelman syndrome 852
- Angina
 stable 744
 unstable 738, 739*f*, 744
- Angiokeratoma corporis diffusum. *See* Fabry disease
- Angiopietin-like protein 3 (ANGPTL3) 721
- Angiotensin-converting-enzyme (ACE)
 28–29, 29*f*
 in cerebrospinal fluid 667
- Angiotensin-converting-enzyme (ACE) inhibitors
 control of pathological thirst 41
 diabetes management
 cardiovascular risk 309
 kidney disease 323
 hyperkalaemia risk 59
 proteinuria management
 in chronic kidney disease 156
 membranous nephropathy 158
 microalbuminuria 164–165
- Angiotensin I 28–29, 29*f*, 127
- Angiotensin II 28–29, 29*f*, 127, 137
 control of thirst 32, 41
 in stress response 406
- Angiotensin II receptor blockers (ARBs)
 diabetes management
 cardiovascular risk 309
 kidney disease 323
 microalbuminuria management 164–165
 proteinuria management
 in chronic kidney disease 156
 membranous nephropathy 158
 microalbuminuria 164–165
- Angiotensinogen 28–29, 29*f*
- Anion gap 74
 in poisoning 875*t*
- Ankylosing spondylitis 638
- Ankyrins 516–517
- Anorchia 431, 455
- Anorexia nervosa 207
- amenorrhoea and 369
 endocrine abnormalities 207, 207*b*, 681
 potassium depletion 53
- Antacid-induced hypophosphataemic osteomalacia 621*t*, 622, 623
- Antenatal screening 4–5. *See also* Fetus, screening for malformation. *specific techniques*
- Anthropometric measurements, nutritional status 193–194
- Antiarrhythmics, therapeutic drug monitoring 777–778
- Antibiotic therapy
 shocked patients 409–411
 therapeutic drug monitoring
 aminoglycoside antibiotics 770–771, 773, 781
 chloramphenicol 781
 glycopeptides 781
- Antibodies
 heterophilic, immunoassay interference 385, 826*b*
 non-pathogenic 583*t*
 pathogenic 582, 583*t*. *See also* Autoantibodies
 specific responses, quantification 602
- Antibody identification panels 511
- Antibody screening 511
- Antibody tests
 chronic CNS infections 669–670
 pernicious anaemia 524
- Anticholinergic syndrome 790*t*
- α_1 -Antichymotrypsin 572*t*
- Anticoagulants, antidotes 794*t*
- Anticonvulsants, therapeutic drug monitoring 778–780
 newer drugs 779–780
- Anti-D antibodies 510, 512–513
- Antidepressants
 poisoning 803–804
 therapeutic drug monitoring 780
- Antidiuretic hormone. *See* Arginine vasopressin (AVP)
- Antidotes 794, 794*t*
- Antiemetics, as cause of hyperprolactinaemia 361
- Antiepileptics. *See* Anticonvulsants
- Antifungal drugs, therapeutic drug monitoring 781–782
- Antigen(s) 561
 blood group 510–511
 distribution, nucleus 587, 587*f*
 presentation 570
 receptors 561
 B lymphocytes. *See* Immunoglobulin(s)
 T lymphocytes. *See* T cell receptor (TCR)
 recognition 561, 566–570
- Antihypertensive therapy 765
 hyperlipidaemia association 731
 sodium excess 39
- Anti-inflammatory drugs, therapeutic drug monitoring 777
- Antimicrobial drugs
 therapeutic drug monitoring 781–782.
See also Antibiotic therapy
- Antimitochondrial antibodies (AMA) 256
- Antimitotic syndrome 790*t*
- Anti-Müllerian hormone (AMH) 413, 434–435, 437, 451–452
 measurement 437
 in disorder of sex development (DSD) 420, 422*t*
 in infertile women 440–441
- Antineoplastic drugs, therapeutic drug monitoring 782–783
- Antineutrophil cytoplasmic antibodies (ANCA) 585, 586*f*
- Antinuclear antibodies (ANAs) 255–256, 585, 586*f*, 587, 587*t*
 in pregnancy 587
- Antinuclear factor (ANF), antibodies against 638, 639*t*, 644–645
- Anti-oestrogen preparations 439
- Antiphospholipid antibodies 587–588
- Antipsychotic drugs
 metabolic complications
 hyperglycaemia 682
 hyperlipidaemia 682, 732
 hyperprolactinaemia 681
 therapeutic drug monitoring 780
- Antiretroviral drugs
 hyperlipidaemia association 732
 therapeutic drug monitoring 782
- Antithrombin 408
- Antithyroid drugs 376
 Graves disease treatment 391
 maternal–fetal transfer 380
 thyroid stimulating hormone-secreting pituitary tumour treatment 393
 toxic multinodular goitre treatment 392
- α_1 -Antitrypsin
 deficiency 254*t*, 255, 268, 495
 DNA analysis 862–863
 gene 862
 location 845
 liver function tests 242
 properties 572*t*
- Antitubercular drugs, therapeutic drug monitoring 782
- Anuria 129
 in acute kidney injury 136
- Anxiety 676–677, 681
- Aplastic anaemia 524
- Apnoea of prematurity 486
- Apolipoprotein 708–710
 characteristics 707*t*
 functions 707*t*, 708
 measurement, investigation of lipid disorders 733
 in nephrotic syndrome 160*t*
- Apolipoprotein(a) 710
 cardiovascular disease risk and 753
- Apolipoprotein A 707*t*, 709, 713
 abnormal structure 722*t*, 728
 A-I 707*t*, 709
 deficiency 722*t*, 728

- A-I Milano 728
 A-II 707*t*, 709
 A-IV 707*t*, 709
 A-V 707*t*, 709
 Apolipoprotein B 707, 707*t*, 709
 measurement, investigation of lipid disorders 733
 Apolipoprotein B-48 224, 707–708, 709, 713
 characteristics 707*t*
 Apolipoprotein B-100 707*t*, 709, 714, 866
 familial defective apolipoprotein B-100 727, 866
 Apolipoprotein B-containing lipoproteins 709
 assembly 703–704, 711–713
 Apolipoprotein C 707*t*, 709–710
 C-I 707*t*, 710
 C-II 707*t*, 710, 713, 714
 deficiency 725
 C-III 707*t*, 710, 713
 Apolipoprotein D 707*t*, 710
 Apolipoprotein E 707*t*, 710
 investigation of lipid disorders 733
 polymorphism 710, 719, 752, 867
 Apolipoprotein M 707*t*, 710
 Appetite 203–205
 central control 203–204
 in cancer cachexia 817
 peripheral signals 204–205
 salt 30
 Aprataxin, gene mutation 700
 APUDomas 226–227
 Aquaporins 27, 30, 129
 Arachidonic acid 705, 706*f*
 Arcuate nucleus, appetite control and 203
 Arginine
 structure 170*f*
 supplementation 410
 Arginine infusion, assessment of growth hormone reserve 356
 Arginine vasopressin (AVP) 31, 129, 352, 404
 in ascites formation 259
 in chronic dilutional hyponatraemia 47, 49*f*
 in chronic kidney disease 147
 collecting duct and 129
 in diabetes insipidus 39, 40–41
 cranial 39–40, 40*f*
 in hypodipsic hypernatraemia 44, 45–46, 45*f*
 non-osmotic control 31
 osmoregulatory control 31, 31*f*
 plasma measurements, pituitary function assessment 354
 in pregnancy 41
 renal responsiveness 31, 32*f*
 secretion 352
 ectopic 814, 815*t*
 fetal 487
 water deprivation test. *See* Water deprivation test
 stress response 404, 405, 406
 Arginine vasopressin (AVP) receptor antagonists 51
 Arginine vasopressin (AVP) V₂ receptor (AVPR2) 31, 32*f*
 in congenital nephrogenic diabetes insipidus 40–41, 43
 Arginyl-glycyl-aspartate (RGD)-containing glycoproteins 607, 607*t*
 Aromatase 438
 Aromatic amino acid decarboxylase (AADC) deficiency 478
 Arteries, structure and function 741, 741*f*
 Arthritis
 Charcot 643
 crystal 639–643
 inflammatory 637–638. *See also* Rheumatoid arthritis (RA)
 osteoarthritis. *See* Osteoarthritis (OA)
 Arthus reaction 575
 Articular cartilage 636–637, 637*f*
 Articular disorders 637–643
 aches and pains 638–639
 autoimmune 585, 601*t*, 638, 639*t*, 644–645
 in endocrine disease 643
 laboratory testing 643–645
 in systemic disease 643. *See also* Arthritis. *specific disorders*
 Articular system 636–637
 disorders. *See* Articular disorders
 Ascites 259–260
 bacterial infections and 264–265, 265*t*
 causes 259–260, 260*b*
 investigations 260, 260*t*, 261*f*
 monitoring 260
 L-Ascorbic acid 189. *See also* Vitamin C
 Asfotase alfa 624
 Asialotransferrin 666, 666*f*
 Aspartate, metabolism 70
 Aspartate aminotransferase (AST)
 in acute hepatitis 251
 in alcoholic hepatitis 257
 liver function tests 240
 mitochondrial isoenzyme 241
 in muscle disease 653
 in pregnancy 446. *See also* Aminotransferases
 Aspirin 705
 management of diabetes-related cardiovascular risk 307–308
 pharmacological and metabolic effects 797*f*
 poisoning 796–797
 clinical features 796
 laboratory measurements 792*t*, 796–797
 management 797
 mechanisms 796, 797*f*
 therapeutic drug monitoring 777
 Asplenic patients, immune deficiency 580
 Assisted reproduction 441, 457. *See also* In vitro fertilization (IVF)
 Asterix 699
 Asymmetrical proximal neuropathy 691
 Ataxia 699–701
 in cerebrotendinous xanthomatosis 700–701
 in ceroid lipofuscinosis 701
 in coeliac disease 701
 early-onset with oculomotor apraxia and hypoalbuminemia 700
 fragile-X-associated tremor/ataxia syndrome 700
 Friedreich 699
 in hexoaminidase deficiency 700
 with vitamin E deficiency 185, 699–700
 Ataxia telangiectasia 700
 Atelectasis 89
 Atherogenesis
 cholesteryl esterase transfer protein and 718
 lipoprotein lipase role 716
 theories of 741–743
 lipid oxidation hypothesis 741–743, 742*f*
 response-to-injury hypothesis 741
 vitamin E and 743
 Atherosclerosis 738, 741–743, 755
 diabetic populations 298
 DNA analysis 866
 eicosanoid role 705
 high density lipoprotein association 715
 lecithin cholesterol acyltransferase and 716
 Atherosclerotic plaque 743, 743*f*
 complicated/rupture 743, 743*f*
 ATP binding cassette (ABC) transporter family 720
 Atrial flutter/fibrillation, in hyperthyroidism 388
 Atrial natriuretic peptide (ANP) 29, 30, 30*f*
 ectopic secretion 815*t*
 in hyperaldosteronism 38–39
 in idiopathic oedema 38
 menstrual cycle and 38
 in pre-eclampsia 38
 in pregnancy 37
 Audit, laboratory services 23, 23*b*, 23*f*
 Autoantibodies 582
 detection 582, 603
 tests, in connective tissue diseases 644–645
 to thyroidal antigens 386–387, 583, 601*t*. *See also* Autoimmune diseases
 Autohaemolysis test 530–531
 Autoimmune diseases 582–588
 articular 585, 601*t*, 638, 639*t*, 644–645
 endocrine 583–588, 601*t*
 gastrointestinal tract 583, 601*t*
 kidneys 585, 585*t*, 586*f*, 601*t*
 liver 584–585, 584*t*, 601*t*
 skin 585, 601*t*. *See also* Autoantibodies
 Autoimmune encephalopathy 688, 688*t*
 Autoimmune haemolytic anaemia 527*b*, 528
 in malignant disease 816
 morphological features 501, 503*f*
 Autoimmune hepatitis (AIH) 255–256, 584
 investigations 584, 584*t*
 monitoring response to therapy 256
 in pregnancy 263
 Autoimmune hypoglycaemia 344
 Autoimmune hypoparathyroidism 106
 Autoimmune insulin syndrome 344
 Autoimmune polyendocrine syndromes (APS) 583–585, 584*t*
 Autoimmune polyendocrinopathy–candidiasis–ectodermal dystrophy (APECED) syndrome 106, 107*t*, 584*t*
 Autoimmune polyglandular syndrome type 1 (APS-1) 106
 Autoimmune polyglandular syndrome type 2 (APS-2) 106
 Automated peritoneal dialysis (APD), chronic kidney disease 150
 Autonomic failure, hypoglycaemia-associated 320
 Autonomic nervous system (ANS), stress response 403–404, 404*f*
 Autonomic neuropathy, diabetic 324
 Autosomal dominant hypocalcaemia with hypercalciuria (ADHH) 106
 Autosomal dominant hypophosphataemic rickets (ADHR) 112, 113
 Autosomal dominant inheritance 467, 467*f*, 850, 850*f*
 Autosomal recessive hypercholesterolaemia (ARH) 727–728
 Autosomal recessive hypophosphataemic rickets (ARHR) 113
 Autosomal recessive inheritance 466–467, 467*f*, 850, 850*f*
 Avidin, biotin deficiency and 189
 Azathioprine therapy 600*t*
 autoimmune hepatitis 256
 Azodipyrrole 239

- B**
- Bacterial flora, commensal 218, 561
- Bacterial infections 597–598, 597*t*
 asplenic patients 580
 diagnosis and monitoring 598
 liver function test abnormalities 264–265
 sickle cell patients 556
 white blood cell morphological abnormalities 504, 505*f*. *See also* Infection(s)
- Bacterial overgrowth, small bowel. *See* Small bowel bacterial overgrowth
- Bacterial septicaemia, hypoglycaemia association 347
- Barbiturate abuse 801*t*
- Bardet–Biedl syndrome 202*t*
- Bariatric surgery 205–206
 combined restrictive and malabsorptive procedures 206, 206*f*
 contraindications 205–206
 monitoring after 206, 207*t*
 neuropathy associated with 694
 resolution of type 2 diabetes mellitus following 206, 286–287, 291, 318
 restrictive procedures 206, 206*f*
 type 2 diabetes management 291, 317–318
- Barker hypothesis 286
- Bartter syndrome 55
 non-respiratory alkalosis 84
- Basal metabolic rate (BMR), calculation 196, 196*f*
- Base deficit 74
- Base excess 73
- Basic calcium phosphate (BCP) crystals 639, 642
- Basic calcium phosphate deposition disease 642
- Basic multicellular unit (BMU) 608
- Basophil(s) 563*t*, 565
 count, reference ranges 498*t*, 563*t*
 morphological features 504*f*
- Basophilic stippling, excessive 529
- Bazedoxifene 618
- B cell malignancies 594–597
 investigation 588. *See also* Monoclonal proteins
 paraprotein types 591, 591*t*
- Becker muscular dystrophy 653
 DNA analysis 864–865, 864*f*
 gene mapping 845*f*
- Beer potomania 49
- Behavioural syndromes 675*t*
- Bence Jones escape 590
- Bence Jones protein (BJP) 588
 in amyloidosis 596
 in B cell malignancy 590
 in heavy chain disease 595
 immunofixation 592, 592*f*
 measurement 590, 593
 in monoclonal gammopathy of unknown significance 596
 in myeloma 594
 in Waldenström macroglobulinaemia 595
- Bence Jones proteinuria 163–164, 163*t*
- Benzodiazepine overdose 803
 antidote 794*t*, 803
- Beri-beri. *See* Thiamin (vitamin B₁) deficiency
- β-Cells. *See* Pancreatic β-Cells
- β-Blockers
 antidotes 794*t*
 Graves disease management 391
 hypertension management, in diabetes mellitus 309
 hypoglycaemia association 346
 redistribution hyperkalaemia and 58
- β-Thalassaemia 552, 553–554, 554*b*, 554*t*, 557–558
- Bias 10–11, 12
- Bicarbonate
 body fluid composition 28*t*
 buffering 66–67
 diabetic ketoacidosis management 328–329
 extrarenal fluid composition 34*t*
 fractional excretion 179
 measurement 73
 non-respiratory acidosis management 76
 reabsorption 70, 71*f*, 82, 169, 169*t*
 renal absorption 128
 renal excretion, non-respiratory alkalosis 82–83
 renal tubular handling 169, 169*t*
- Bicarbonate, plasma concentration
 analytical goals 11*t*
 analytical variation 11*t*
 biological variation 11*t*
 critical difference 15*t*
- Biglycan 606–607
- Biguanides 313. *See also* Metformin
- Bile 233–234, 711
 fluid composition 34*t*
 secretion 236
- Bile acid(s)
 enterohepatic circulation 236, 236*f*, 711
 liver function tests 243
 metabolism 711
 plasma
 in intrahepatic cholestasis of pregnancy 262–263
 measurement 243
 sequestrants 735*t*
 synthesis 236, 236*f*
 defects 495, 711
- Bile ducts 234
 tests of patency 247
- Bile pigments, metabolism 237–239
- Bile salts 224, 236
- Biliary atresias 494–495
- Biliary canaliculi 233*f*, 234
- Biliary cirrhosis, primary 256
- Biliary drainage 234
- Biliary tract disease, neoplastic 264
- Biliary tract obstruction
 alkaline phosphatase activity 239. *See also* Cholestasis
- Biliopancreatic diversion 206, 206*f*, 207*t*
- Bilirubin
 analytical goals 11*t*
 analytical variation 11*t*
 biological variation 11*t*
 in cerebrospinal fluid 661, 666
 detection 667
 conjugated 237, 238*f*, 493
 in haemolysis 529
 quantitation 239
 critical difference 15*t*
 enterohepatic circulation 237–238, 238*f*
 liver function tests 237, 247
 plasma concentrations 238, 238*f*
 in chronic hepatitis 254
 elevated. *See* Hyperbilirubinaemia
 in pre-hepatic jaundice 245
 quantitation 239
 unconjugated 237, 493
 in haemolysis 525, 529
 quantitation 239
 urinary 131, 245. *See also* Bilirubinuria
- Bilirubin metabolism 237–239, 237*f*, 238*f*
 disorders
 inherited, presenting in childhood 494. *See also* Hyperbilirubinaemia
 neonates 493
- Bilirubinuria 245
 in acute hepatitis 251
- Biliverdin 237, 237*f*
- Bioavailability 768
- Biochemical data 6–20
 acquisition 6, 7*t*
 comparison. *See* Comparison of results
 factors affecting 7–13
 analytical 10–12
 postanalytical 12–13
 preanalytical 7–10
 interpretation 13–16
 predictive value. *See* Predictive value of tests
 uses 1–5. *See also* Tests
- Biological factors, affecting test results 7–10, 8*t*
- Biological variation, intrinsic 9–10, 11*t*, 15
- Biologics, rheumatoid arthritis treatment 638
- Biopterin 477
 metabolism defects 670
- Biothesiometry, investigation of impotence 459
- Biotin 189
 deficiency 189
 laboratory-based assessment 198
- Biotransformation, hepatic 235
- Bipolar affective disorder 678
- Birth weight, low 484
 as cardiovascular disease risk factor 752
- Bisphosphonate therapy
 acute phase response 617–618
 hypercalcaemia 105
 osteogenesis imperfecta 633–634
 osteoporosis 616, 617–619
 biochemical responses 619
 Paget disease, biochemical responses 631–632, 631*f*
- Bitot's spots 185
- Black tea 757
- Blackwater fever 528
- Bladder
 cancer, tumour markers 827
 urinary diversion from 54
- Bleeding
 as cause of anaemia 524
 in vitamin K deficiency 186–187. *See also* Haemorrhage
- Blood 515
 clotting. *See* entries beginning *coagulation*
 count. *See* Full blood count (FBC)
 donation 510
 glucose. *See* Glucose, blood
 grouping 511, 511*t*
 ketone testing 299–300
 oxygen uptake. *See* Oxygen uptake
 peripheral, as stem cell source 600
 pH 66
 products 513
 transfusion. *See* Blood transfusion;
 Haematology
- Blood–brain barrier (BBB)
 appetite control and 203
 permeability
 assessment 664
 proteins 662–663
- Blood cell morphology 501–507
 in haematological malignancies 504–507
 red blood cells. *See* Red blood cell morphology
 white blood cells. *See* White blood cells (WBCs)
- Blood film 501
- Blood film examination 501
- Blood flow, renal. *See* Renal blood flow
- Blood gas analysers 73, 88, 91
- Blood gas analysis
 inherited metabolic disease 469
 psychiatric disorder investigation 676*t*

- Blood group antigens 510–511
 Blood pressure measurement 757
 penile 459
 Blood transfusion 510–514
 antibody identification panels 511
 β thalassaemia management 553
 blood products 513
 crossmatching 511–512
 haemolytic reaction, investigation 512
 laboratory tests 511–512
 regulations 514
 risks 513
 sickle cell disease 556
 universal donors 510
 universal recipients 510
 Blood vessels
 arteries, structure and function 741, 741f
 renal 127
 B lymphocytes 563t, 566, 588–594
 antigen presentation 570
 antigen receptors. *See* Immunoglobulin(s)
 clonality 561
 deficiency
 associated infections 576, 577t
 primary immunodeficiencies 577–578
 Body fluid distribution 27–28, 28t
 Body mass, effect on test results 8
 Body mass index (BMI) 193, 193t, 201–202
 relationship with type 2 diabetes
 risk 286, 286f
 Body temperature, management in poisoning 793
 Bohr effect 76, 550–551
 Bone(s) 604–613
 age-related changes 609, 615
 anatomy 604–605
 macroscopic 604
 microscopic 604–605
 biopsy, indications 617, 620, 635
 chronic kidney disease–mineral and bone disorder 627
 calcium fluxes 94f, 95
 cellular elements 605, 607–609
 cortical (compact) 604
 disease. *See* Bone disease
 effect of acidosis 76
 flat 604
 formation, markers 609–611, 610f
 hyperthyroidism manifestations 389
 lamellar 604–605
 long 604
 loss, age-related 609
 matrix proteins 605–607
 pain, pagetic 630
 parathyroid hormone receptors 96
 remodelling 95, 607–609, 609f
 resorption 608, 609f
 markers 611–613
 roles 604
 trabecular 604
 turnover
 assessment of calcium metabolism 100
 in children 632, 632t
 in Paget disease 630
 turnover markers 609–613, 610b
 in chronic kidney disease–mineral and bone disorder 626
 new 613
 in osteoporosis 616, 616t
 in Paget disease 630–631
 variation in 613
 Bone densitometry 615–616
 chronic kidney disease–mineral and bone disorder investigation 627
 critical values 749
 factors affecting 749
 Buccal mucosa 215
 Buffering 66–67, 68
 ammonia 67
 bicarbonate 66–67
 haemoglobin 67, 67f
 non-respiratory acidosis 74
 non-respiratory alkalosis 82
 phosphate 67
 plasma proteins 67
 respiratory acidosis 80–81
 respiratory alkalosis 84
 Bulimia nervosa 207, 208, 681
 Bullous porphyrias 542–545
 biochemical features and diagnostic approach 536f, 537t, 543
 congenital erythropoietic porphyria. *See* Congenital erythropoietic porphyria (CEP)
 hepatoerythropoietic porphyria 542–543, 545
 pathophysiology 543
 porphyria cutanea tarda. *See* Porphyria cutanea tarda (PCT)
 skin symptoms and signs 543, 543f
 Buprenorphine, therapeutic drug monitoring 785
 Butane abuse 803
 γ -Butyrolactone (GBL) 876–877
- C**
 C1 esterase inhibitor 571t
 deficiency 579
 C282Y mutation 265
 Cabergoline therapy
 acromegaly 363
 hyperprolactinaemia 361, 440
 Cachexia, cancer. *See* Cancer cachexia
 Cadmium, tubular damage 161
 Caecum, microflora 218
 Caerulein 217
 Caeruloplasmin 190, 198–199, 243
 liver function tests 243
 properties 572t
 in Wilson disease 243, 266
 Caffeine
 effect on levothyroxine absorption 381–382
 therapeutic drug monitoring 783
 Calcidiol. *See* 25-Hydroxyvitamin D (25(OH)D)
 Calciferol 622. *See also* Vitamin D
 Calcification 114
 Calcimimetic drugs 104–105, 628
 Calcineurin inhibitors 150
 Calcinosis, tumoral 113–114
 Calciopenic osteomalacia. *See* Osteomalacia
 Calciphylaxis 114
 Calcipotriol 98
 Calcitonin 98–99
 ectopic secretion 815t
 effect on TmP/GFR 112
 as marker for medullary thyroid carcinoma 400, 812, 823t
 osteoporosis management 619
 Paget disease management 632
 Calcitonin gene-related peptide, ectopic secretion 815t
 Calcitriol. *See* 1,25-Dihydroxyvitamin D (1,25(OH)₂D)
 Calcitriol therapy
 hypophosphataemic osteomalacia 623
 osteoporosis 617
 vitamin D-dependent rickets 622

- Calcium
 absorption test 100, 121–122, 617
 interpretation 122
 analysis of tubular handling 100, 104, 122
 interpretation 122, 122*f*
 biological role 93
 body fluid composition 28*t*
 deficiency 93
 dietary intake 94
 bone health and 615
 osteoporosis management 617
 postmenopausal women 617
 distribution 93, 94*t*
 effect on TmP/GFR microalbuminuria 112
 endogenous faecal 94
 fluxes 94–95, 94*f*
 intestinal absorption 94, 94*f*
 increased 175
 measurement 100
 test. *See absorption test (above)*
 plasma concentration. *See Calcium, plasma concentration*
 stones 175–177, 175*t*
 supplementation, osteoporosis management 617
 urinary excretion 95
 urinary measurement 100
 in osteoporosis 616
 in Paget disease 631
- Calcium, plasma concentration 93, 99
 analytical goals 11*t*
 analytical variation 11*t*
 biological variation 11*t*
 in chronic kidney disease 144
 in chronic kidney disease–mineral and bone disorder 625–627, 626*b*
 critical difference 15*t*
 diurnal variation 100, 100*f*
 in hypoparathyroidism 106
 measurement 99–100
 in osteopenia of prematurity 492, 492*t*
 in Paget disease 631
 in poisoning 875*t*
- Calcium-based phosphate binders 114–115, 148–149
- Calcium channel blockers, hypertension management in diabetes mellitus 309
- Calcium gluconate 61, 109, 121
- Calcium metabolism 93–109
 biochemical assessment 99–100
 bone 94*f*, 95
 in chronic kidney disease 144
 gastrointestinal tract 94, 94*f*
 kidney 94–95, 94*f*
 in neonates. *See Neonates*
 regulation 95–99
- Calcium pyrophosphate (CPP) crystals 639
- Calcium pyrophosphate deposition (CPPD) 642
- Calcium sensing receptor (CaSR) 102
 mutations 491
- Calculi, renal. *See Renal calculi*
- Calorimetry 195
- Calprotectin 225
 in disease 225–226
- Campesterol 757
- Campylobacter jejuni*, food contamination 190
- Cancer
 breast. *See Breast cancer*
 cachexia. *See Cancer cachexia*
 cervical. *See Cervical cancer*
 colorectal. *See Colorectal cancer*
 diet and 208, 208*t*
 protective effect of dietary fibre 192, 208
 gastric. *See Gastric cancer*
 gene therapy 871
 genetics 867–869
 liver, primary. *See Hepatocellular carcinoma (HCC)*
 lung, tumour markers 836–837, 836*t*
 pancreatic, tumour marker 840
 parathyroid gland 101, 102*t*
 prostate. *See Prostate cancer*
 testicular. *See Testicular cancer*
 thyroid gland. *See Thyroid cancer; See also Malignancy; Tumour markers*
 Cancer antigen 15-3 (CA 15-3) 823*t*, 826*t*
 in breast cancer monitoring 828–829
 Cancer antigen 19-9 (CA 19-9)
 as cancer marker 823*t*, 826*t*
 gastric cancer 831
 pancreatic cancer 840
 in primary sclerosing cholangitis 256–257
 Cancer antigen 125 (CA125) 822*t*, 826*t*
 factors interfering with levels 838, 838*t*
 as ovarian cancer marker 824*t*, 838
 detection of residual disease 839
 diagnosis 839
 long-term surveillance 840
 monitoring 839–840
 prognosis 839
 screening 838–839
 Cancer cachexia 816–818
 metabolic changes 817–818, 817*f*, 817*t*
 treatment 818
 Cancers of unknown primary origin, tumour markers 842
 Cannabis abuse 801*t*, 802–803
 Capillary zone electrophoresis (CZE), paraprotein identification/quantitation 589–590, 592
 Capsulitis, shoulder 643
 Carbamazepine
 measurement, indications 792*t*
 therapeutic drug monitoring 778
 Carbidopa 477
 Carbimazole 391
 Carbohydrate(s)
 absorption 220–222, 221*f*
 clinical aspects 222
 investigation 222
 deficiency 182
 dietary 181–182, 220–221
 digestion 221–222
 enteral feeding 210
 intake in diabetes 306
 parenteral nutrition 211
 Carbohydrate-deficient transferrin (CDT) 242–243, 258
 Carbohydrate metabolism
 in chronic kidney disease 148, 148*b*
 disorders of 654
 investigation 654–655
 in malignancy 818
 muscle pain and 650–651. *See also specific disorders*
 hepatic role 234
 in neonates 488–490, 489*f*
 Carbonate dehydratase 67, 67*f*, 71*f*, 72*f*
 Carbon dioxide
 buffering 66
 erythrocytes 67, 67*f*
 daily production and elimination 68, 68*t*, 73
 excretion 68*t*, 70, 80
 hydrogen ion production 68, 68*t*
 partial pressure. *See PCO₂*
 retention 80. *See also Acidosis, respiratory*
 total, measurement 73
 Carbonic acid 66, 68, 70
 Carbonic anhydrase II deficiency 634
 Carbon monoxide 805
 Carbon monoxide poisoning 685, 805
 clinical features 805
 management 794*t*, 805
 mechanisms 805
 pregnant women 788
 Carboxyhaemoglobin 788–791, 805
 measurement 789–791, 792*t*, 805
 Carboxyl ester lipase 717
 Carcinoembryonic antigen (CEA) 264, 822*t*, 824*t*, 826*t*
 as colorectal cancer marker
 diagnosis 830
 monitoring 830–831
 prognosis 830
 as gastric cancer marker 831
 as lung cancer marker 836*t*, 837
 as medullary thyroid cancer marker 812
 Carcinoid crisis 808–809
 Carcinoid syndrome 226
 clinical features 808–809, 809*b*
 Carcinoid tumours 808–810
 clinical presentation 808–809, 809*b*
 diagnostic imaging 810
 foregut 811
 intestinal 226
 laboratory investigation 809–810
 serotonin metabolism 809
 treatment 810
 Cardiac failure. *See Heart failure*
 Cardiac glycosides, therapeutic drug monitoring 777–778
 Cardiac muscle 740–741, 740*f*
 creatine kinase 652. *See also Creatine kinase-MB (CK-MB)*
 Cardiac troponins. *See Troponin(s)*
 Cardiorespiratory failure, encephalopathy and 687
 Cardiovascular disease (CVD) 737–738
 diabetes-related risk, pharmacological management 307–309
 effect of hormonal contraceptives on risk 448
 hyperlipidaemia-associated risk, reduction 734
 microalbuminuria as marker of risk 164–165
 mortality 737–738, 738*f*
 risk assessment 750–751, 751*f*
 risk factors. *See Cardiovascular risk factors; See also Atherosclerosis*
 Cardiovascular risk factors 750–757
 dietary 208, 756–757
 potentially modifiable 752–756
 relative importance of 755–756
 unmodifiable 751–752
 Cardiovascular system 737
 arteries, structure and function 741, 741*f*
 cardiac muscle 740–741, 740*f*
 disease. *See Cardiovascular disease (CVD)*
 functions 737
 Carney complex 813
 Carnitine 189
 measurement in plasma, tissue and urine 657
 therapy 482–483
 transport, measurement 657–658
 Carnitine deficiency 657–658
 Carnitine palmitoyltransferase, measurement of activity 657–658
 Carnitine palmitoyltransferase I (CPT I) 488–489, 489*f*, 647–649, 649*f*
 β-Carotene 184, 184*f*, 185
 β-Carotene therapy, erythropoietic protoporphyria 547
 Carotenoids 184

- Carpal tunnel syndrome
 in acromegaly 693
 in hyperthyroidism 693
 in hypothyroidism 332, 693
- Carpenter syndrome 202*t*
- Carrier 850
 definition 871
- Cartilage, articular 636–637, 637*f*
- Cartilaginous joints 636
- Casts, urinary 131
- Catecholamines
 in chronic kidney disease 147
 effect on potassium distribution 32
 measurement, pheochromocytoma
 diagnosis 763–764, 763*t*
 factors affecting results 763–764, 764*t*
 in post-traumatic stress disorder 679
 in redistribution hypokalaemia 52
 stress response 9, 405, 407–408
 tumour marker characteristics 823*t*
- Cathepsin K 613, 619
 inhibitors 619
- Cation exchange resins 54
- Caveolae 703–704
- CD (cluster of differentiation) 562–564,
 565*t*
- CD36 719–720, 742
- CD117 antigen 831–832
- CD ligand deficiency 578
- Cell counting, blood cells 498
- Cell-mediated immunity
 primary immunodeficiencies 578. *See also*
 T lymphocytes
- Cell membranes, cholesterol and 703–704
- Cellulose 192
- Cellulose acetate electrophoresis,
 haemoglobinopathy investigation 557
- Central cyanosis 88
- Central nervous system (CNS)
 acidosis effects 76
 acute infections 669
 acute porphyria manifestations 538–539
 chronic infections 669–670
 disorders
 biochemical investigations 669–671
 inflammatory 701
 non-biochemical investigations
 668–669, 668*b*
 hypercapnia effects 81
 hyperthyroidism manifestations 389
 hypothyroidism manifestations 394–395
 insulin actions 280*t*
 poisoning-related complications 791–793
- Central pontine myelinolysis (CPM)
 51, 687
- Ceramide 705, 707*f*
- Ceramide trihexosidase deficiency. *See* Fabry
 disease
- Cerebral oedema, in diabetic ketoacidosis
 329
- Cerebral salt wasting 50
- Cerebrosides 705
- Cerebrospinal fluid (CSF) 660–672
 in acute CNS infections 669
 appearance 661
 cells 661–662
 in chronic CNS infections 669–670
 cisternal 661
 composition 661
 in dementia 670–671, 688
 in demyelinating diseases 671
 enzymes 667
 ferritin 666–667
 formation 661
 glucose 662
 in meningitis 669
- haem pigments 661, 666–667
 detection 667
- inflammation markers 667–668
- in inherited metabolic diseases 670
- in intracranial haemorrhage 670
- investigations 661–669
 biochemical, in central nervous system
 disorders 669–671
 non-biochemical 668–669, 668*b*
- lactate 662
 in congenital lactic acidosis 463
 in meningitis 669
 in respiratory chain dysfunction 655,
 662
- in malignancy 670
- obstructed flow 670
- otorrhoea 666
- physiology 661
- proteins 662–665, 663*t*, 671–672
 age-related changes 663, 663*t*
 in assessment of blood–brain barrier
 permeability and reduced fluid flow
 664
 brain-specific 663*f*, 665–666
 increased concentrations, causes 663
 in meningitis 669
 plasma-derived 662–663, 663*f*
 protein index 664
 reduced flow 663–664
 assessment 664
 rhinorrhoea 666, 666*f*
 sampling and pressure 661
 ventricular 661
- Cerebrotendinous xanthomatosis 695–696,
 700–701, 711
- Cerebrovascular disease 737–738. *See also*
 Stroke
- Ceroid lipofuscinosis, neuronal 701
- Cervical cancer 829
 monitoring 829
 prognosis 829
 screening and diagnosis 829
- α -Chain disease 595–596
- γ -Chain disease 596
- μ -Chain disease 596
- Chain of custody documents 876, 876*b*
- Charcoal, activated 794, 797, 798, 803
- Charcot foot 324–325
- Charcot joint 643
- Chase–Aurbach test 122
- Cheiroarthropathy, diabetic 643
- Chelation therapy
 iron poisoning 794*t*, 798
 lead poisoning 794*t*, 799
- Chemokines 573*t*
- Chemotaxins 571*t*
- Chemotherapy
 effect on somatic growth 818–819
 hyperkalaemia and 58–59
 malignant insulinoma 342–343
- Chenodeoxycholic acid 236, 236*f*, 711
- Chest pain 744
 tests 748–749. *See also* Myocardial damage,
 acute
- Children
 bone disease 632, 632*t*
 bone turnover 632
 diabetic, nutrition and growth 296
 hypercalcaemia, causes 104*b*
 liver disease 495–496
 conjugated hyperbilirubinaemia and
 hepatocellular disease 494–495, 494*b*
 malignant disease therapies, effects on
 growth 632
 poisoning 788
 protein–energy malnutrition 200, 201
- thyroid function 379
 tumour markers 837–838. *See also* Infants
- Child–Turcotte–Pugh classification 248,
 248*t*
- Chloramphenicol, therapeutic drug
 monitoring 781
- Chloride
 body fluid composition 28*t*
 depletion 55, 57
 extrarenal fluid composition 34*t*
 plasma concentration
 in poisoning 875*t*
 renal tubular acidosis 172
 retention 57
 vitreous humour concentration,
 post-mortem 879, 879*t*
- Chloride-losing diarrhoea, congenital 55
- Chloroquine poisoning 792*t*, 798
- Chloroquine therapy, porphyria cutanea
 tarda 544
- Chlorpromazine 682
- Chloruresis 57
- Cholangitis, primary sclerosing 256–257
- Cholecalciferol 96
- Cholecystokinin (CCK)
 appetite control 204
 in bulimia nervosa 208
- Cholestanolosis 695–696, 700–701, 711
- Cholestasis 236, 245
 alkaline phosphatase activity 239
 lipoprotein X in 708, 731
 in pregnancy 262–263, 445
- Cholestatic jaundice 246
- Cholesterol 703–704
 absorption 710–711
 crystals, synovial fluid 643
 dietary, effect on plasma concentrations
 757
 efflux 715
 function 703*t*
 HDL. *See* High density lipoprotein (HDL)
 cholesterol
 hepatic metabolism 236, 236*f*, 703, 711
 hepatic trafficking 714
 LDL. *See* Low density lipoprotein (LDL)
 cholesterol
 measurement, investigation of lipid
 disorders 732
 in membranes 703–704
 non-HDL, measurement 733
 nuclear 706
 plasma concentration. *See* Cholesterol,
 plasma concentration
 reverse transport 714*f*, 715
 side chain cleavage, ovarian steroidogenesis
 437, 438*f*
 structure 437*f*, 703*f*
 synthesis, defective in Smith–Lemli–Opitz
 syndrome 462
 total 732
 cardiovascular disease risk and 752
- Cholesterol, plasma concentration
 age-related changes 8
 analytical goals 11*t*
 analytical variation 11*t*
 biological variation 11*t*
 critical difference 15*t*
 effect of dietary cholesterol 757
 gender differences 8
 in nephrotic syndrome 160, 160*t*
 normal range 13, 14*f*
 in type 2 diabetes 290
- Cholesterol desmolase deficiency 35,
 353*t*. *See also* Congenital adrenal
 hyperplasia (CAH)
- Cholesteryl ester storage disease 717

- Cholesteryl ester transfer protein (CETP)
709, 714*f*, 715, 718
deficiency 722*t*, 729
inhibitors 735*t*
- Cholic acid 236, 236*f*, 711
- Choline 189
- Cholinergic syndrome 790*t*
- Cholinesterase, measurement 792*t*
- Chondrocalcinosis 642
in haemochromatosis 643
- Chorea 699
- Choriocarcinoma 835. *See also* Gestational trophoblastic neoplasia (GTN)
- Chorionic villus sampling (CVS) 444–445, 444*f*, 444*t*, 476, 860
mutation analysis and 476
- Choroid plexus 661, 662–663
- Chromium 191–192
deficiency 191–192
dietary sources 191–192
laboratory-based assessment 199
toxicity 192, 199, 799
- Chromogranin A (CgA) 809–810, 811, 823*t*
post-mortem biochemistry 879*t*
- Chronic dilutional hyponatraemia 46, 47–49, 47*f*, 49*f*, 50
arginine vasopressin responses to hypertonic saline infusion 47, 49*f*
causes 47, 48*b*
laboratory investigation 50–51
management 51. *See also* Hyponatraemia
- Chronic disease, anaemia of 524, 643–644, 816, 816*t*
- Chronic energy deficiency (CED) 200
in western adults 201
- Chronic granulomatous disease 579
- Chronic inflammatory demyelinating polyneuropathy (CIDP) 692–693
- Chronic kidney disease (CKD) 141–150
acute kidney injury *vs.* 136
acute-on-chronic kidney disease 139, 139*b*
aetiology and pathogenesis 141–142, 141*b*
anaemia 143, 147, 149
calcification 114
carbohydrate metabolism 148, 148*b*
definition, using estimated glomerular filtration rate 134, 135*t*
encephalopathy and 686
endocrine disturbances 144, 145*b*
endocrine control of salt and water balance 147, 147*f*
major abnormalities 144, 145*b*
mechanisms 145*t*
hypertension in 142, 759
hypertriglyceridaemia in 730–731
hypoglycaemia in, predisposing factors 337–338
incidental factors contributing to 142, 142*b*
lipid metabolism 148, 730–731
management 148–150
dietary protein restriction 148–149, 323
general 148–149
prevention of complications 149
renal replacement treatment 149–150
metabolic disturbances 143–144
acidosis 78–79, 143
peripheral neuropathy and 693
prognosis, proteinuria quantitation 156
sexual dysfunction 145–146
sodium excretion 34, 147, 147*f*
staging 135, 135*t*, 148
proteinuria quantitation 156
thyroid abnormalities 146–147, 146*b*
uraemic syndrome. *See* Uraemic syndrome
- Chronic Kidney Disease Epidemiology Collaboration group (CKD-EPI) 134
- Chronic kidney disease–mineral and bone disorder (CKD-MBD) 143, 144, 624–628, 624*f*
acidosis role in pathogenesis 76, 78–79, 625
aetiology 624–625
clinical features 625
investigations 625–627
treatment 627–628
- Chronic liver disease. *See* Liver disease
- Chronic lymphocytic leukaemia (CLL) 505–506, 506*f*, 591*t*, 595–596
- Chronic lymphoid leukaemias 505–506
- Chronic myeloid leukaemia 505, 506*f*
- Chronic pain syndrome, in acute porphyria 539
- Chronic rejection 599*b*
- Chronic sensorimotor neuropathy, diabetic 323–324
- Chronic tophaceous gout 641
- Chylomicron(s) 224, 707–708, 708*t*, 713
modification 713
remnants 707–708, 713
- Chylomicronaemia syndrome 716, 724–725
- Chylomicron retention disease 712–713, 722*t*, 724
- Chymase, post-mortem biochemistry 879*t*
- Chymotrypsin 223
faecal test 218
- Ciclosporin
statin interaction 653
therapeutic drug monitoring 772–773, 784
- Ciclosporin A 600*t*
- Cider potomania 49
- Cinacalcet 104–105, 628
- Circulatory system. *See* Cardiovascular system
- Circumference measurements 194
- Cirrhosis 77, 259–263
ascites 259–260
differential diagnosis 260, 260*t*, 261*f*
monitoring 260
glucose intolerance 263, 295
hepatic encephalopathy 259
Indian childhood 267*t*, 268
in men 262, 262*t*
primary biliary 256
renal impairment 260–261
differential diagnosis 261*t*
sex hormones and their binding proteins 261–263
vascular disturbance 259. *See also* Alcoholic liver disease
- Cisterna magna, cerebrospinal fluid 661
- Citalopram
overdose 804
therapeutic drug monitoring 780
- Citrate–glucuronic acid solutions 121
- Classical pathway, complement activation 570, 571*f*
- Clearance 131, 768–769
creatinine. *See* Creatinine clearance
inulin 132, 132*t*
- Clearance tests, hepatic function 243–244
- Clinical audit, laboratory services 23, 25
- Clinical effectiveness, laboratory medicine 24
- Clinical geneticists, role in disorder of sex development (DSD) investigations/interventions 422–423
- Clinical pseudohypoglycaemia 320
- Clinical quality indicators, laboratory medicine 24–25, 24*t*
- Clitoromegaly 417, 419, 425
- Clomifene 440
- Clomifene test 356–357, 454, 460
- Clomipramine, therapeutic drug monitoring 780
- Clonality 561
- Clone, definition 872
- Clonidine test, assessment of growth hormone reserve 356
- Cloning 854
positional, definition 872
- Clozapine 682
- Cluster of differentiation (CD) 562–564, 565*t*. *See also* entries beginning CD
- Coagulation cascade 507, 508*f*
role of vitamin K 186
- Coagulation factor(s) 507, 508*f*
abnormalities in nephrotic syndrome 159
cardiovascular disease risk and 753
concentrates 513
deficiency of individual factors, detection 509
liver function tests 242
stress response 408–409
- Coagulation tests 507–509
interpretation 509–510, 509*t*
- Coatmer protein 711–712
- Cobalamin. *See* Vitamin B₁₂
- Cobalamin C disorder 476–477
- Cobalt toxicity 799
- Cocaine abuse 801*t*, 802
- Cocaine- and amphetamine-regulated transcript (CART) 203, 204, 811
- Cockcroft and Gault formula 134
- Codeine abuse 801*t*
- Codominant, definition 850, 872
- Codons 846
- Coefficients of variation (CV) 10, 12
- Coeliac disease 220, 583
as cause of iron deficiency anaemia 519–520
diabetes association 296
diagnosis 583
neurological complications 701
therapeutic diets 209*t*
- Coenzyme Q10 209
- Cofactor supplementation, inherited metabolic disease management 479, 479*t*
- Coffee consumption 757
- Colchicine 641
- Colipase 224
- Collagen 605
cross-links 611–612, 612*f*
metabolites, serum tests for hepatic fibrosis 244
synthesis 605, 606*f*
telopeptides 612–613, 612*f*, 630, 632
- Collecting duct 125, 125*f*, 126
function 129
- Colloid osmotic pressure 28
- Colon
fluid and electrolyte absorption 54–55
microflora 218. *See also* Gastrointestinal tract
- Colony-stimulating factors 573*t*
- Colorectal cancer 829
advanced disease, monitoring 831
'deleted in colonic carcinoma' gene (*DCC*) 868
diagnosis 830
faecal calprotectin concentrations 225–226
monitoring 830–831
prognosis and staging 830, 830*t*
protective effect of dietary fibre 192
screening 829–830
DNA analysis 869
tumour markers 829–831
cautions and caveats 831

- Coma
 diabetic 326–327, 687
 Glasgow Coma Scale (GCS) 684, 684*t*
 hyperglycaemic 687
 myxoedema 397
- Combined oral contraceptives 447, 447*t*
 contraindications 448
 metabolic effects 448–449
- Commensal organisms 218, 561
- Common variable immunodeficiency (CVID) 577
- Comparison of results
 observed results *vs.* reference limits 15
 with previous values 15–16, 15*t*
- Compatibility testing (crossmatching) 511–512
- Complement 570–571, 571*f*
 activation pathways 570, 571*f*
 alternative pathway 570, 571*f*
 classical pathways 570, 571*f*
 lectin pathway 570, 571*f*
 biological actions 570, 571*t*
 deficiency 579
 associated infections 576, 577*t*
 quantification 602
 regulation 570–571, 571*t*
- Complementary DNA (cDNA) 853–854
 definition 872
- Complementation studies, inherited
 metabolic diseases 475
- Complement C3 572*t*
- Complement regulatory protein C4b binding protein (C4bBP) 408–409
- Complete androgen insensitivity syndrome (CAIS) 426*t*, 427–428, 456
- Complete gonadal dysgenesis (Swyer syndrome) 425
- Compliance, drug. *See* Adherence, drug
- Compound heterozygotes 848, 851
 definition 872
- Compulsive water drinking 41, 681
- Computed tomography (CT)
 acute kidney injury 138
 pituitary gland 360. *See also* Imaging techniques
- Conception 436
- Confusional state, acute. *See* Delirium
- Congenital adrenal hyperplasia (CAH) 60, 352, 370, 417, 423
 diagnosis 370, 419, 441, 450
 hirsutism 441
 hyperkalaemia 60
 hypertensive forms 761*t*
 hypoglycaemia 345
 hypokalaemic alkalosis 56
 lipoid 426*t*
 sodium wasting 35, 60
 subtypes 353*t*
 treatment 370
- Congenital adrenal hypoplasia 430, 455
- Congenital analbuminaemia 37
- Congenital chloride-losing diarrhoea 55
- Congenital disorder of glycosylation (CDG) 477
- Congenital erythropoietic porphyria (CEP) 534*t*, 545
 biochemical findings 537*t*
 molecular genetics 537*t*, 545
- Congenital nephrogenic diabetes insipidus 40–41, 43
- Congenital septo-optic dysplasia 39–40, 367
- Congestive cardiac failure 37, 749. *See also* Heart failure
- Connective tissue diseases (CTDs) 585, 601*t*, 638, 639*t*, 644–645
- Conn syndrome, secondary diabetes 294
- Continuous ambulatory peritoneal dialysis (CAPD)
 chronic kidney disease 150
 effect on plasma renin activity 147
 effect on thyroid hormones 146
- Continuous positive airways pressure (CPAP) 92
- Continuous subcutaneous insulin infusion (CSII) 311, 312*t*
- Continuous venovenous haemodiafiltration (CVVHDF) 141
- Continuous venovenous haemofiltration (CVVH) 141
- Contraception
 hormonal methods 447, 447*t*
 injectable 447*t*, 449
 oral. *See* Oral contraceptives (OCs)
 non-hormonal methods 447
- Convulsions
 alkalosis and 83
 poisoning-related 791–793. *See also* Seizures
- Coombs test 530
- Co-oximeters 91
- Copeptin 748
- Copper 190–191
 accumulation. *See* Wilson disease
 deficiency 190–191, 198–199, 689, 694
 dietary 190
 laboratory-based assessment 198–199
 overload 191
 plasma concentration, in Wilson disease 266–267
 retention, in primary sclerosing cholangitis 256–257
 tissue concentration, in Wilson disease 267
- Coproporphyrin, hereditary. *See* Hereditary coproporphyrin (HCP)
- Coproporphyrin 537*t*, 539
 urinary
 in Dubin–Johnson syndrome 246, 548
 increased 548, 548*t*
- Coproporphyrinogen oxidase 533–534, 535*f*
- Cord blood, as stem cell source 600*t*
- Cordocentesis 444*t*, 445
- Coronary heart disease 737–738
 risk
 hormone replacement therapy and 449–450
 plasma cholesterol concentration and 13, 14*f*. *See also* Cardiovascular disease (CVD)
- Corrin ring 521, 523*f*
- Corrosive poisoning 790*t*
- Cortical bone 604
- Corticosterone methyl oxidase deficiency 35, 59
- Corticotrophin. *See* Adrenocorticotrophic hormone (ACTH)
- Corticotrophin releasing factor (CRF) 350–351
- Corticotrophin releasing hormone (CRH) 350–351
 deficiency 369
 ectopic secretion 815*t*
 labour role 446, 447
 stress response 404, 404*f*, 405
 tests 356
 Cushing syndrome 365, 372
- Cortisol
 anti-inflammatory effects 407
 anti-insulin effect 405, 407
 in chronic kidney disease 147
 in Cushing disease 365
 in depression 678–679, 680
- diurnal variation 8–9
 in hyperthyroidism 389
 in hypoglycaemia-associated autonomic failure 320
 in hypothyroidism 395
 midnight salivary 364
 pituitary function assessment 357, 358
 assay precision 355
 basal hormonal investigations 353–354
 borderline responses 355
 dynamic test reproducibility 355
 insulin stress test 354
 normal ranges 355
 short tetracosactide test 354
 steroid therapy and 358
 in post-traumatic stress disorder 679, 680
 secretion 350–351
 hypoglycaemia response 335
 stress response 9, 405, 407
 urinary free, Cushing syndrome screening 363
- Cortisol-binding globulin (CBG) 405, 438–439
- C-peptide
 insulin biosynthesis 278–279
 plasma concentration, measurement 303
 forensic biochemistry 877, 877*b*
 hypoglycaemia investigation 337*f*, 338–339
 in type 1 diabetes mellitus 284, 285
- Cranial diabetes insipidus (CDI) 39
 causes 39–40, 40*b*
 hypertonic saline infusion 42
 management 42
 water deprivation test 42
- Craniopharyngiomas 367
- C-reactive protein (CRP) 598
 analytical range 10
 cardiovascular disease risk and 755
 cerebrospinal fluid 667–668
 in acute CNS infections 669
 measurement, clinical applications 598, 598*t*, 644
 microalbuminuria and 164
 post-mortem biochemistry 879*t*, 882
 properties 572*t*
 in rheumatoid arthritis 585, 644
- Creatine kinase (CK) 646–647, 652, 747
 analytical goals 11*t*
 analytical variation 11*t*
 biological variation 11*t*
 critical difference 15*t*
 effect of exercise 9, 652
 ethnic origin differences 8
 increased activity, causes 9, 652, 652*b*
 statins 653
 plasma measurement, muscle disease investigation 651–653
- Creatine kinase-BB (CK-BB) 652
 in cerebrospinal fluid 667
- Creatine kinase-MB (CK-MB) 652, 747
 as biomarker of acute myocardial damage 744, 745*f*, 747–748
- Creatine kinase-MM (CK-MM) 652
- Creatinine 132
 plasma concentration. *See* Creatinine, plasma concentration
 post-mortem biochemistry 879*t*
 renal tubular handling 169*t*
 urinary, assessment of calcium metabolism 100
- Creatinine, plasma concentration
 age-related changes 8
 analytical goals 11*t*
 analytical variation 11*t*
 biological variation 11*t*

- Creatinine, plasma concentration (*Continued*)
 in chronic kidney disease 143
 critical difference 15*t*
 effect of food intake 9
 interpretation of results 15
 in obese patients 8
 relationship with creatinine clearance
 132–133, 133*f*
 renal function assessment 132–133, 133*f*
 Creatinine clearance 132, 132*t*
 calculation 134
 in pregnancy 445
 relationship with plasma creatinine
 concentration 132–133, 133*f*
 Creutzfeldt–Jakob disease (CJD) 670–671
 Crigler–Najjar syndrome 246, 494
 Critical difference 15, 15*t*
 Critically ill patients, definitions 409, 409*b*
 Crohn disease 220
 faecal calprotectin concentrations
 225–226. *See also* Inflammatory bowel
 disease (IBD)
 Crossmatching 511–512
 Crossover 849, 849*f*
 definition 872
 Cryofibrinogen 593, 593*b*
 Cryofibrinogenaemia 593
 Cryoglobulins 593, 593*b*
 Cryoprecipitate 513
 Cryoproteins 593–594, 593*b*
 Cryptorchidism 455
 Crystal arthritis 639–643
 Crystals, urine 131
 C-terminal telopeptide (CTX) assay 612*f*,
 613, 630
 C-type natriuretic peptide (CNP) 30
 Cubilin 154, 521
 Cushing syndrome
 ACTH-dependent. *See*
 Adrenocorticotrophic hormone
 (ACTH)-dependent Cushing
 syndrome
 clinical features 363
 depression and 678–679
 diagnosis and differential diagnosis
 363–366, 364*f*, 441
 corticotrophin releasing hormone test
 365, 372
 imaging 366
 management 366
 obesity and 203
 reassessment after pituitary surgery 366
 secondary diabetes 293, 294
 subclinical 370
 Cutaneous porphyrias 534*t*, 542–547
 bullous. *See* Bullous porphyrias
 erythropoietic protoporphyria. *See*
 Erythropoietic protoporphyria (EPP)
 X-linked dominant protoporphyria. *See*
 X-linked dominant protoporphyria
 (XLDPP)
 Cyanide–nitroprusside test 472
 Cyanide poisoning 805
 antidote 794*t*, 805
 Cyanosis 88
 Cyclic adenosine monophosphate (cAMP),
 parathyroid hormone and 95–96
 Cyclical oedema 38
 Cyclic citrullinated peptide (CCP),
 antibodies to 586, 644
 Cyclophilin B 605
 CYFRA 21-1 823*t*, 836*t*, 837
 CYP2D6 776–777, 862
 Cyproterone acetate 440
 Cystathionine β -synthase deficiency,
 pyridoxine challenge 479
 Cystatin C 135
 cerebrospinal fluid 663*f*, 663*t*, 666
 Cysteamine 483
 Cysteine 183
 structure 171*f*
 Cystic fibrosis (CF)
 DNA analysis 863–864, 864*f*
 hepatobiliary complications 268–269
 hypokalaemia 55
 molecular therapies 481
 screening 217–218
 sweat test 863
 Cystic fibrosis transmembrane conductance
 regulator (CFTR) 863
 gene mutations 863
 Cystine 170*f*
 stones 170–171, 175*t*, 177
 structure 171*f*
 urinary excretion, normal 171
 Cystinosis 174, 483, 492
 Cystinuria 170–171
 clinical manifestation 170–171
 mode of inheritance 171
 treatment 171
 Cytochrome *c*, reduced 656
 Cytochrome *c* oxidase (COX) 655, 655*f*
 Cytochrome P450
 paracetamol metabolism 795, 795*f*
 pharmacogenetic studies 776–777, 862
 Cytochrome P450 aromatase deficiency 423,
 424*t*
 Cytochrome P450 oxidoreductase deficiency
 (ORD) 420, 423, 424*t*, 425, 426*t*
 Cytokines 572–575
 endocrine hormones *vs.* 572, 574*t*
 inflammatory 566, 574
 major groups 406, 573*t*
 stress response 406
 Cytokine therapy, as cause of hypothyroidism
 397
D
 Data transfer errors 13
 Dawn phenomenon 320–321
 D-dimer measurement 508–509, 509*t*,
 748–749
 Decontamination, intestinal 794
 Decorin 606–607
 Deferasirox 553–554
 Deferiprone 553–554
 Dehydration, brain cell response 45
 Dehydroepiandrosterone (DHEA)
 in females 439
 production, in puberty 414
 Dehydroepiandrosterone sulphate (DHEAS)
 in females 439
 elevated, in hirsutism 441
 production, in puberty 414
 Delayed puberty 368–369, 429–430
 boys 420, 430
 causes 430*b*
 delayed growth and 369, 430
 endocrine investigations 428*t*
 Delayed-type hypersensitivity reactions 575
 'Deleted in colonic carcinoma' gene (*DCC*)
 868
 Delirium 676
 causes 676, 677*t*
 Delta checks 16
 Demand management, laboratories 25
 Demeclocycline, nephrogenic diabetes
 insipidus and 41
 Dementia 677–678, 688
 cerebrospinal fluid analysis 670–671, 688
 Demyelinating diseases 671
 Denaturation, DNA 852
 Denaturing high performance liquid
 chromatography (DHPLC), mutation
 screening 857
 Denosumab 619
 Densitometry, bone. *See* Bone densitometry
 Dental caries 208
 Dent disease 172
 Deoxyypyridinoline 611–612, 612*f*
 Deoxyribonucleic acid. *See* DNA
 (deoxyribonucleic acid)
 Deoxyuridine suppression test 524
 Depression 678–679
 adrenal function and 678–679, 680
 features 678
 growth hormone secretion abnormalities
 681
 in metabolic syndrome and diabetes 679
 thyroid function and 678, 680
 Depressive adjustment disorder 678
 De Quervain thyroiditis 399
 Desamino-8-D-arginine vasopressin. *See*
 Desmopressin (dDAVP)
 Desferrioxamine 544, 553–554, 628
 iron overdose treatment 794*t*, 798
 test, dialysis patients 626–627, 635
 17,20-Desmolase 437–438
 Desmopressin (dDAVP) 41
 polyuria management 42
 replacement therapy 371
 stimulation test 364
 water deprivation test 42, 357
 DEXA (dual-energy X-ray absorptiometry)
 220, 615–616
 Dexamethasone suppression test
 high-dose 365
 low-dose 363–364
 overnight 364
 depressed patients 679
 Dextromethorphan 482
 Dextrose saline 36*t*
 Diabetes insipidus (DI) 39, 354, 360, 369,
 488
 after pituitary surgery 359
 causes 39–40, 40*b*
 cranial. *See* Cranial diabetes insipidus
 (CDI)
 investigations 354
 hypertonic saline infusion 357
 water deprivation test. *See* Water
 deprivation test
 nephrogenic. *See* Nephrogenic diabetes
 insipidus (NDI)
 polyuria 39–40, 40*f*
 water deficiency 44
 Diabetes mellitus 273–274
 alcohol-related/pancreatic causes 293
 articular disorders 643
 biochemical measurements 299–303
 β -Cell function measurement 303
 glucose measurements 283*b*, 283*t*, 299
 hyperinsulinaemic clamps 303
 insulin resistance 301–303
 intravenous glucose tolerance testing
 (IVGTT) 303
 ketone testing 299–300, 322
 oral glucose tolerance test. *See* Oral
 glucose tolerance test (OGTT)
 post-mortem 878*b*, 880, 881*f*
 recent glycaemic control 300–301
 as cardiovascular disease risk factor 754,
 754*f*
 classification 282–295, 285*b*
 complications 322–326
 acute 322
 chronic 322–323

- effect of intensive glycaemic control 319
 macrovascular 296, 297–298, 307
 mechanisms of diabetic tissue damage 296–298
 microvascular 296, 297
 pregnancy-related 331*b*. *See also specific complications*
 definitions 282–283
 depression and 679
 diagnosis 282–295, 299
 diagnostic criteria 282, 283*b*, 283*t*, 284*b*
 emotional stress and 322
 endocrine associations 295–296
 feet in 324–325
 gestational. *See Gestational diabetes mellitus*
 growth 296
 high-risk individuals 301
 homeostasis model assessment (HOMA) 303
 hypertriglyceridaemia in 729–730
 hypoglycaemia in 339–341
 classification 339–340, 340*b*
 definition 339–340
 incidence 340
 management 341
 pathophysiology and risk factors 340, 340*b*
 tissue damage 297
 hypophosphatemia 116
 iatrogenic 294
 impotence and 458
 inadequately controlled 298–299
 insulin resistance. *See Insulin resistance*
 management 302*f*, 305–332
 brittle diabetes 326
 cardiovascular risk, pharmacological management 307–309
 complications 322–326. *See also specific complications*
 diet 306
 exercise 306–307
 general aspects 305–309
 glucose-lowering therapy. *See Glucose-lowering therapy, diabetes*
 in hospital setting 330–331
 hypoglycaemia 340–341
 intercurrent illness 321–322
 obstacles to achieving glycaemic control 318–322
 patient education 307
 in pregnancy 331–332
 'sick day rules' 321–322
 smoking cessation 307
 therapeutic diets 209*t*
 type 1 diabetes. *see under Diabetes mellitus, type 1*
 type 2 diabetes. *see under Diabetes mellitus, type 2*
 maturity onset diabetes of the young. *See Maturity onset diabetes of the young (MODY)*
 mechanisms of tissue damage 296–298
 microalbuminuria 164
 nutrition 296, 306
 oral contraceptives and 448–449
 pathogenesis 296–297
 polyuria 39
 post-mortem biochemistry 878*b*, 880, 881*f*
 research investigations 303
 role of intracellular metabolic pathways for glucose 277–278
 screening 301, 302*f*
 secondary 285*b*, 287, 293–295
 tropical 293–295
 Diabetes mellitus, type 1 283–286, 583
 aetiology 284–286
 complications
 effect of intensive glycaemic control 319
 microalbuminuria 322
 nephropathy 322
 retinopathy, screening 325. *See also Diabetes mellitus, complications*
 dyslipidaemia and 290
 environmental cofactors 284–286
 genetic associations 286, 286*t*
 honeymoon period 283–285, 310
 hypertension and 289
 incidence 282
 latent autoimmune diabetes of adults (LADA) 283–284
 management
 immunotherapy 318
 insulin use 310–313. *See also Insulin therapy*
 intercurrent illness 322. *See also Diabetes mellitus, management*
 microalbuminuria 164
 oral contraceptives and 448–449
 patient education 307
 pregnancy and 291–292, 331–332
 Diabetes mellitus, type 2 286–291
 abnormalities of non-insulin-mediated glucose disposal 276, 289
 accelerator hypothesis 283–284
 amylin role 288
 associations 289–290
 dyslipidaemia 290, 307
 hypertension 289–290
 metabolic syndrome 289–290, 679
 obesity 286, 286*f*, 289
 complications
 effect of intensive glycaemic control 319
 microalbuminuria 322
 nephropathy 322
 retinopathy, screening 325. *See also Diabetes mellitus, complications*
 genetic factors 287
 glucoregulatory defects 287–289, 287*t*
 glucose transporters in 276, 289
 heterogeneity 286–287
 insulin-requiring/insulin-treated patients 288
 insulin resistance 286–287, 288
 lipotoxicity and glucotoxicity 290–291
 management
 bariatric surgery 291, 317–318
 dietary 306
 exercise 307
 insulin 313, 317
 intensive glycaemic control 319
 intercurrent illness 321
 liver disease and 263
 metformin. *See Metformin*
 patient education 307
 maturity onset diabetes of the young. *See Maturity onset diabetes of the young (MODY)*
 microalbuminuria 164
 natural history 289, 289*f*
 oral contraceptives and 448–449
 pancreatic β -Cell deficiency/dysfunction 286–288
 pregnancy and 331–332
 prevalence 282
 prevention studies 291
 resolution following bariatric surgery 206, 286–287, 291
 Diabetic amyotrophy 324
 Diabetic coma 326–327, 687
 Diabetic hand syndrome 643
 Diabetic ketoacidosis (DKA) 69, 76, 299, 326–329
 alcoholic ketoacidosis *vs.* 326–327, 330
 post-mortem biochemistry 878*b*, 880, 881*f*
 biochemical features 327, 327*b*
 ketone testing 299, 327
 management 327–329
 bicarbonate 328–329
 cerebral oedema 329
 fluids 328
 general measures 327–328
 insulin 328
 potassium, magnesium and phosphate replacement 328
 mechanisms 326, 327*f*
 normoglycaemic 326–327
 potassium depletion 55, 328
 resolution 329
 Diabetic nephropathy 322
 end-stage disease 323
 management 323
 microalbuminuria as marker of risk 164, 322–323
 Diabetic neuropathy 323–324, 643, 690*t*, 691–692
 pathophysiology 691–692
 presentations 691, 691*b*
 focal and multifocal neuropathies 691
 symmetrical polyneuropathies 691
 Diabetic retinopathy 325–326
 screening 325, 325*b*
 in pregnancy 331
 Diagnosis 2–3. *See also specific diseases*
 Diagnostic odds ratio (DOR) 20
 Dialysis
 inherited metabolic disease management 483. *See also Haemodialysis; Peritoneal dialysis*
 Dialysis-treated patients
 desferrioxamine test 626–627, 635
 hypercalcaemia 625–626, 626*b*
 hyperphosphataemia 626, 626*b*
 sexual dysfunction 146
 Diamond–Blackfan syndrome 524
 Diamorphine. *See Heroin*
 Diaphysis 604
 Diarrhoea
 congenital chloride-losing 55
 magnesium-related 120–121
 potassium depletion 53–54
 Diazoxide 314, 342
 Dicarboxylic acids, measurement in urine 657
 DIDMOAD syndrome 39–40
 Diet(s)
 in aetiology of disease 208
 cancer and. *see under Malignancy*
 cardiovascular risk factors 208, 756–757
 dental caries and 208
 effect on urinary protein excretion 155
 ketogenic 478, 478*f*
 therapeutic 208–209, 209*t*. *See also Dietary management*
 Dietary assessment 193
 Dietary management
 chronic kidney disease 148
 diabetes mellitus 306
 hypercalcaemia 176
 hyperlipidaemia 209*t*, 734
 obesity 205
 renal stones 178–179
 calcium-containing 176. *See also Nutritional management*
 Dietary pattern, cardiovascular disease risk and 757
 Dietary reference value (DRV), definition 181*t*
 Dietary supplements 208–209

- Diethylenetriaminepentaacetic acid (DTPA), radiolabelled, glomerular filtration rate measurement 135–136
- Differentiated thyroid cancer (DTC) 399
- DiGeorge syndrome 491, 578
- Digitoxin, therapeutic drug monitoring 777–778
- Digoxin
poisoning 792*t*, 798
clinical features 798
management 794*t*, 798
therapeutic drug monitoring 773, 773*f*, 774, 777–778
- Digoxin-like interfering substances (DLIS) 778
- Dihydrocodeine abuse 801*t*
- Dihydrofolate 522*f*
- Dihydropteridine reductase deficiency 477
- Dihydrotestosterone (DHT) 262, 452
biological actions 452*t*, 456
effect on hair follicles 441
in fetal sex development 413
human chorionic gonadotrophin stimulation test 422
- 1,25-Dihydroxyvitamin D (1,25(OH)₂D) 94–95, 97
action
defects 621
sites of 98
cancer and 98
in chronic kidney disease 144
in chronic kidney disease–mineral and bone disorder (CKD–MBD) 624–625
inactivation 97
measurement 98, 197, 621–622, 621*t*
in osteomalacia 620–621, 623
phosphate regulation 110
synthesis
defects 621
fetal 490
sites 97
synthetic analogues 98. *See also* Calcitriol therapy
- Di-iodotyrosine (DIT) 376
- Dilutional acidosis 78
- 2,3-Dimercaptopropane sulfonate (DMPS) 799
- 2,3-Dimercaptosuccinic acid (DMSA) 799
- Dipeptidyl peptidase IV (DDP-4) inhibitors 317
adverse effects 317
hypoglycaemia 317, 341
- 2,3-Diphosphoglycerate (2,3-DPG) 550
concentration, effect of changes on oxygen–haemoglobin dissociation curve 90, 90*f*
effect of acidosis 76
effect of alkalosis 82
effect of phosphate deficiency 110
measurement 531
production 517, 518*f*
- Diploid, definition 872
- Direct antiglobulin test 530
- Direct calorimetry 195
- Disaccharidases 221, 221*f*
- Disaccharides 220–221, 221*f*
intestinal hydrolysis, differential tests 222, 223*f*
- Disease management 3–4. *See also specific diseases*
- Disease modifying anti-rheumatic drugs (DMARDs) 638
- Disease progression, monitoring 3–4. *See also specific diseases*
- Disease severity, assessment 3
- Disopyramide
hypoglycaemia association 346
therapeutic drug monitoring 778
- Disorder of sex development (DSD) 412, 416–428
classification 423–428, 424*t*
definition 416
evaluation of external genitalia 417–418, 418*f*
evaluation of internal anatomy 418
general examination of newborns with suspected DSD 417
investigations
adolescents with DSD 419–420, 419*f*
anti-Müllerian hormone 420, 422*t*
human chorionic gonadotrophin stimulation test 422, 423*f*
inhibins 421–422
insulin-like factor 3 420
newborns with DSD 418–419
role of clinical geneticists 422–423
steroid measurement and its interpretation 420, 421*f*
management, general principles 417
terminology 416, 417*t*
XX DSD. *See* XX disorders of sex development
XY DSD. *See* XY disorders of sex development. *See also specific disorders*
- Disseminated intravascular coagulation (DIC) 501–503, 503*f*, 509–510
- Distal convoluted tubule (DCT) 125, 125*f*, 126
function 129
tests 136
magnesium homeostasis 118
- Diuresis 129
primary renal sodium loss 34
- Diuretic(s)
abuse 35
acute dilutional hyponatraemia management 51
chloride depletion 55
hyperlipidaemia association 731
hypertension management in diabetes mellitus 309
potassium-sparing, potassium retention 59
thiazide. *See* Thiazide diuretics
- Diurnal variation
aldosterone:renin ratio 762
bone turnover markers 613
calcium concentration 100, 100*f*
effect on test results 8–9
parathyroid hormone 95
phosphate concentration 100*f*, 113
urinary protein excretion 155
- DNA (deoxyribonucleic acid) 844, 845, 846
chips 856–857
denaturation 852
detection of specific sequences 852–855
double-stranded, antibodies to 587
methylation 847, 851–852
mitochondrial. *See* Mitochondrial DNA (mtDNA)
probes 852–855
recombinant 852, 853*f*
repair 847
defects 868–869
restriction enzymes. *See* Restriction endonuclease(s)
sequencing 857–858, 857*f*
next generation 859–860. *See also entries beginning gene/genetic*
- DNA analysis
applications 860–869
cancer genetics 867–869
diagnosis of index cases 860
haemoglobinopathy investigation 558
multifactorial and polygenic disease 866–867
prenatal diagnosis 860–861
screening 861
single gene disorders 862
techniques. *See* Genetic analysis, techniques
- Döhle bodies 504, 505*f*
- Dominant, definition 872
- Dopamine
control of prolactin secretion 351, 681
inhibition of thyroid-stimulating hormone release 379, 379*f*
- Dopamine agonist therapy
acromegaly 363
monitoring response 362
prolactinomas 361
side effects 362
- Dopamine infusion test, acromegaly diagnosis 362
- Dopamine receptor antagonists 681
- Dopa-responsive dystonia (DRD) 698
- Dosulepin, therapeutic drug monitoring 780
- Double-stranded DNA, antibodies to 587, 638, 644–645
- Downstream, definition 872
- Down syndrome 678
screening 443–444, 444*b*, 861
- Doxepin, therapeutic drug monitoring 780
- Doxercalciferol 98, 627
- D-penicillamine, cystinuria treatment 171, 171*f*
- Drinking 31–32
- Drowning, post-mortem biochemistry 880
- Drug(s)
absorption 768, 768*f*
adherence. *See* Adherence, drug
affecting aldosterone:renin ratio 762, 762*b*
affecting catecholamine and metadrenaline measurements 763–764, 764*t*
affecting thyroid function 381–382, 381*t*
distribution 768, 768*f*
effect on test results 8*t*, 9
effect on TmP/GFR 112, 112*t*
elimination 768–769, 768*f*
half-life 768*f*, 769, 773, 773*t*
indications for measurement in suspected/actual poisoning 792*t*. *See also* Poisoning
inherited metabolic disease management 482–483
interactions. *See* Drug interactions
protein binding 769
safe prescribing, in acute porphyria 541
steady-state concentrations 773, 773*t*
dose requirements, calculations 774–775, 786
effect of loading doses 774–775, 775*f*
therapeutic index 770
toxicity, monitoring 771
urine 131. *See also* Therapeutic drug monitoring (TDM). *specific drugs*
- Drug abuse 800–803
maternal 485
screen 792*t*
- Drug-induced conditions
acute pancreatitis 229*t*
amiodarone-induced thyroid dysfunction 390*t*, 393
antacid-induced hypophosphataemic osteomalacia 621*t*, 622, 623
aplastic anaemia 524
delirium 676, 677*t*
diabetes mellitus 294
erectile impotence 459*b*

- gynaecomastia 458t
 hyperlipidaemia. *See* Hyperlipidaemia
 hyperprolactinaemia 361
 hypoglycaemia. *See* Hypoglycaemia
 hypolipidaemia 732
 liver damage 254–256, 263t
 statin-induced elevation of creatine kinase activity 653
 tubular damage 161
 Drug interactions 771, 772t
 statins 653, 736
 Drug-metabolising enzymes, polymorphisms 861
 Dry reagent sticks 299
 DTPA (diethylenetriaminepentaacetic acid), radiolabelled, glomerular filtration rate measurement 135–136
 Dual-energy X-ray absorptiometry (DEXA) 220, 615–616
 Dual oxidase (DUOX) 376
 Dubin–Johnson syndrome 246, 548
 in pregnancy 263
 Duchenne muscular dystrophy 650–651
 creatine kinase activity 652
 diagnosis 653
 DNA analysis 864–865, 864f, 865f
 gene mapping 845f
 Dukes staging system 830, 830t
 Duodenal ulcers, *Helicobacter pylori* infection and 215
 Duodenum
 calcium absorption 94
 iron absorption 519
 phosphate absorption 111. *See also* Gastrointestinal tract
 Dupuytren contracture 643
 Dynamic function tests 63–64
 Dysalphalipoproteinaemias 722t, 728
 Dysbetalipoproteinaemias 722t
 decreased beta lipoproteins. *See* Hypobetalipoproteinaemia
 increased beta lipoproteins 722t. *See also specific disorders*
 Dysgerminomas 838
 Dyskinesia 697
 Dyslipidaemia
 cardiovascular disease risk and 753
 in chronic kidney disease 142, 149
 classification 722t
 diabetic 290, 298, 307, 308
 management 308. *See also* Hyperlipidaemia. *specific disorders*
 Dyslipoproteinaemias
 drug therapy 735t
 primary 722t, 723–729. *See also specific disorders*
 Dyspepsia 214
 Dyspnoea
 in heart failure 749
 in hyperthyroidism 389
 Dysrhythmias, hypomagnesaemia and 119
 Dysthymia 678
 Dystonia 698
 Dystrophin 864
 Dystrophin gene mutations, analysis 864–865, 864f, 865f
 DYT1 dystonia 698
- E**
 Eating disorders 207–208. *See also specific disorders*
 'Ecstasy' 801–802, 801t
 Ectopic hormone secretion. *See* Hormone(s)
 Ectopic pregnancy 229–230
 diagnosis 442
 Ectopic thyroid tissue 390t, 393
 EDTA (ethylenediaminetetraacetic acid), radiolabelled, glomerular filtration rate measurement 135–136
 Education
 continuing, therapeutic drug monitoring staff 776
 diabetes 307
 Effective osmolality 27
 Efficiency of test 16–17, 17t
 Ehlers–Danlos syndrome 632–633, 633t
 Eicosanoids 705
 function 703t, 705
 synthesis 705, 706f
 Eicosapentaenoic acid 705
 Eicosatrienoic acid 705
 Elastase 223
 faecal test 218
 Elastic lamina 741, 741f
 Elderly
 hyperthyroidism 389
 renin secretion 762
 testosterone deficiency 457
 thyroid function 379
 Electrolyte abnormalities, parenteral nutrition-associated 211
 Electrolytes
 in acute kidney injury 140
 body fluid composition 27, 28t
 requirements, infants *vs.* adults 487t
 urine, sodium deficiency 35. *See also specific electrolytes*
 Electronic crossmatch 511–512
 Electronic laboratory handbook 23
 Electronic patient records, integration of laboratory information systems 23
 Electron transport flavoprotein (ETF) 647–649, 649f
 measurement of activity 657–658
 Electrophoresis
 haemoglobin, haemoglobinopathy investigation 557, 558f
 lipoprotein 734
 mutation detection 857, 857f
 paraprotein identification 589, 590f, 591f
 Elimination rate constant 769
 Elimination techniques, poisoning 795
 Elliptocytosis, hereditary 526
 Ellsworth–Howard test, modified 109, 122
 Emergencies, in diabetes mellitus 326–330
 Emotional stress. *See* Stress
 Encephalopathy 683–688
 autoimmune 688, 688t
 causes 684, 684b, 684t
 clinical examination 684
 definition 683–684
 hepatic 259, 686, 686t
 laboratory investigations 685, 685t
 septic 687–688
 toxic and metabolic 685–687
 uraemic 686
 Wernicke 187, 686
 'Encode' project 845–846
 Endocrine disorders
 in anorexia nervosa 207, 207b
 articular system and 643
 autoimmune 583–588, 601t
 in bulimia nervosa 208
 in chronic kidney disease. *See* Chronic kidney disease (CKD)
 diabetes-associated 295–296
 hypertension and 759, 760t
 hypoglycaemia due to 345
 peripheral neuropathy and 693
 in post-traumatic stress disorder 679
 in psychiatric disease 680–681
 in secondary diabetes 293–294. *See also specific endocrine abnormalities*
 Endocrine-disrupting chemicals 453
 Endocrine tumours
 pancreas 226–227
 secondary diabetes 294. *See also specific endocrine tumours*
 Endogenous factors, affecting test results 8, 8t
 Endogenous faecal calcium 94
 Endoscopic retrograde cholangiopancreatography (ERCP) 218, 245, 256–257
 Endoscopy 214
 Endothelial dysfunction, potential causes 741, 741b
 Endothelial lipase (EL) 716–717
 Energy
 deficiency disorders 463
 deficient intake
 hypoglycaemia due to 347. *See also* Protein–energy malnutrition (PEM)
 expenditure, laboratory-based assessment 195–196, 195f, 196f
 metabolism, in red blood cells 517, 518f
 requirements 181–183
 carbohydrate 181–182
 fat 182–183
 Enhancer, definition 872
 Enoyl-CoA hydratase 649f
 measurement of activity 657–658
 Enteral feeding 210
 in severe sepsis/septic shock 410
 Enterocytes
 carbohydrate digestion and absorption 221–222, 221f
 cholesterol absorption 711, 713
 fat absorption 224, 711, 713
 vitamin B₁₂ absorption 521
 Enterohepatic circulation
 bile acids 236, 236f, 711
 bilirubin and urobilinogen 237–238, 238f
 Enterokinase 223
 Enteropathy, protein-losing 224
 Environment, laboratory 22
 Enzyme(s)
 analysis, in inherited metabolic disease 475
 drug-metabolising, polymorphisms 861
 plasma, liver function tests 239–241
 red blood cell defects. *See* Red blood cell(s)
 replacement therapy 478–479
 restriction. *See* Restriction endonuclease(s)
 serum, pancreatic function testing 217–218. *See also specific enzymes*
 Enzyme-linked immunosorbent assay (ELISA), investigation of chronic CNS infections 669–670
 Eosin-5 maleimide (EMA) binding test 531
 Eosinophil(s) 563t, 565
 count, reference ranges 498t, 563t
 morphological features 504f
 Ephrin (Eph) family of signalling proteins 608–609
 Epidermal growth factor receptor gene (EGFR), mutation analysis in lung cancer 837
 Epigenetic, definition 872
 Epinephrine. *See* Adrenaline
 Epiphysis 604
 Epstein–Barr virus (EBV) infection, diagnosis 500–501
 Equilibrium dialysis 383
 Erectile impotence. *See* Impotence
 Ergocalciferol 96
 Errors, analytical 16
 sources 7–13

- Erythrocytes. *See* Red blood cell(s)
 Erythrocyte sedimentation rate (ESR) 499
 as marker of acute phase response 598, 644
 Erythrocytosis, in malignant disease 816
 Erythropoiesis 516*f*, 517
 iron requirements 519, 519*t*
 Erythropoietic protoporphyria (EPP) 534*t*,
 546–547
 biochemical findings 537*t*
 chronic complications and their
 management 547
 molecular genetics 537*t*
 skin symptoms and signs 546–547, 546*f*
 treatment 547
 Erythropoietin
 in chronic kidney disease 147
 ectopic secretion 815*t*
 recombinant 149
 Escitalopram, therapeutic drug monitoring
 780
 Essential amino acids 183, 183*b*
 Essential fatty acids (EFAs) 182–183, 183*f*
 deficiency 183. *See also* Fatty acid(s)
 Essential hypernatraemia 44
 Essential hypertension 758–759, 761*t*. *See also*
 Hypertension
 Established renal failure (ERF), in diabetes
 mellitus 323
 Estimated average requirement (EAR),
 definition 181*t*
 Ethambutol, therapeutic drug monitoring 782
 Ethanol. *See* Alcohol
 Ethinylestradiol 439, 447–448
 Ethnic origin, effect on test results 8
 Ethosuximide, therapeutic drug monitoring
 778
 Ethylenediaminetetraacetic acid (EDTA),
 radiolabelled, glomerular filtration
 rate measurement 135–136
 Ethylene glycol, poisoning 685, 792*t*, 801*f*
 antidotes 794*t*
 hyperoxaluria and 176, 800
 Ethyl glucuronide (EtG) 258
 Ethyl sulphate (EtS) 258
 Etidronate 617–618, 632
 European liver fibrosis score (ELF) 244–245
 Evaluation, laboratory services 23, 23*b*, 23*f*
 Evidence-based clinical biochemistry 25
 Exenatide 316
 Exercise
 in diabetes mellitus 306–307, 321
 effect on bone density 615
 effect on cardiovascular disease risk 754
 effect on creatine kinase activity 9, 652
 effect on test results 9
 hypoglycaemia and 321, 340
 hypophosphataemia and 116
 weight loss and 205
 Exercise-induced proteinuria 155–156
 Exercise-related amenorrhoea 369
 Exercise test, forearm 654, 659
 Exercise testing, assessment of growth
 hormone reserve 356
 Exogenous factors, affecting test results 8–9, 8*t*
 Exome, definition 872
 Exon 845
 definition 872
 skipping 481
 Expansion acidosis 78
 Expression analysis 871
 Expressivity, definition 872
 External genitalia
 ambiguous 456
 evaluation 417–418, 418*f*
 neonates. *See* Neonates; *See also*
 Disorder of sex development (DSD)
 females, development 413–414
 males, development 413
 External masculinization score (EMS) 417,
 418*f*
 External quality assurance (EQA) 21
 Tumour marker measurements 826
 Extracellular fluid (ECF) 27–28, 28*t*
 Expansion, dilutional acidosis 78
 potassium 28*t*, 32–33
 sodium 28–30, 28*t*, 128
 volume reduction calculation 35, 62
 Extractable nuclear antigens (ENAs),
 antibodies to 587, 587*t*
 Extrapontine myelinolysis 51
 Extrarenal fluids 34, 34*t*
 Extrarenal sodium loss 34, 34*b*
 Eye disease, diabetic. *See* Diabetic
 retinopathy
 Eye examinations, diabetic retinopathy
 screening 325
 Eyes, in Graves disease. *See* Graves disease,
 ophthalmology
 Ezetimibe 711, 735*t*
- F**
 Fabry disease 465, 695
 manifestations 695
 Facioscapulohumeral muscular dystrophy
 653
 Factitious hypoglycaemia 341
 Factor V 507, 508*f*
 liver function test 242
 Faecal tests
 fat excretion 225
 intestinal inflammation 225–227
 occult blood 226, 829–830
 pancreatic function 218
 False negatives 16, 16*t*, 17*t*
 False positives 16, 16*t*, 17*t*
 Familial amyloid polyneuropathy (FAP)
 696
 Familial benign hyperphosphatasia 610
 Familial combined hyperlipidaemia (FCH)
 713–714, 722*t*, 724, 752
 Familial defective apolipoprotein B-100
 (FDB) 727
 Familial expansile osteolysis 633*t*, 634
 Familial hyperaldosteronism 55–56
 Familial hypercholesterolaemia (FH) 722*t*,
 726–728, 752
 autosomal recessive 722*t*, 727–728
 classic 722*t*, 726–727
 diagnostic criteria 727*b*
 genetic testing 734, 866–867
 tendon xanthomata 726, 726*f*, 727*f*
 Familial hyperproinsulinaemia 279, 280*t*
 Familial hypertriglyceridaemia 713–714,
 722*t*, 724–725
 Familial hypobetalipoproteinaemia (FHBL)
 722*t*, 724
 Familial hypocalcaemic hypercalcaemia
 (FHH) 102, 102*t*, 103–104, 121, 491
 Familial hypokalaemic periodic paralysis
 (FHPP) 53
 Familial lipoprotein lipase inhibitor 725
 Familial pseudohyperkalaemia 58, 60
 Familial renal iminoglycinuria 172
 Family history, cardiovascular disease risk
 factors 752
 Family studies 541, 541*t*
 Fanconi anaemia 524
 Fanconi syndrome 80, 174, 622*b*
 causes 174, 174*b*
 Farnesoid X-activated receptor (FXR) 720
 Fast, diagnostic 474–475
 Fasting states, glucose homeostasis 333–334,
 334*f*
 Fat
 absorption 224–225
 clinical aspects 224–225
 investigation 225
 dietary 182–183, 224
 digestion 224, 711
 faecal excretion 225. *See also*
 Triglyceride(s)
 Fat-soluble vitamins 184–187, 192, 224
 functions 703*t*. *See also specific vitamins*
 Fatty acid(s) 704
 cis 704*f*
 dietary, cardiovascular disease risk and
 756
 essential. *See* Essential fatty acids (EFAs)
 free. *See* Free fatty acids
 function 703*t*
 metabolism 234–235. *See also* Fatty acid
 oxidation
 non-esterified. *See* Free fatty acids
 saturated 704, 704*f*
 structure 704*f*
 trans 189–190, 704*f*
 transport into mitochondrial matrix
 647–649, 649*f*
 unsaturated 704, 704*f*
 very long chain. *See* Very long chain fatty
 acids (VLCFAs)
 Fatty acid-binding protein (FABP),
 heart-type 745*f*, 748
 Fatty acid oxidation 647–649, 649*f*, 657, 704
 defects 656–657
 investigation 656–658
 muscle pain and 650–651
 flux measurements 657
 screening 461
 Fatty acid transport proteins (FATPs) 718
 Fatty liver. *See* Steatosis
 Fatty streak 742, 743, 743*f*
 Favism 526
 Febuxostat 642
 Fed states, glucose homeostasis 333–334,
 334*f*
 Feet, in diabetes mellitus 324–325
 Felonious hypoglycaemia 341
 Female(s)
 puberty
 endocrinology 414
 physical signs 415, 415*f*
 precocious 428
 reproductive function 433–450
 hormones regulating 436–437
 physiology 433–436
 reproductive steroid hormones. *See*
 Ovarian steroid hormones; *See also*
 Menstrual cycle; Pregnancy
 sex development 412, 413–414, 413*f*,
 849–850
 sexual dysfunction, in chronic kidney
 disease 146
 Female infertility 440–441
 in chronic kidney disease 146
 investigations 440
 Ferricyanide 656
 Ferrihaemate 162–163
 Ferritin 519
 in cerebrospinal fluid 666–667
 in haemochromatosis 265–266
 measurement 520
 in rheumatoid arthritis 644
 Ferrochelatase (FECH) 533–534, 535*f*
 decreased activity 546
 Ferroportin 519, 520*f*
 gene mutations 265

- Fertile eunuch syndrome 455
- Fetal haemoglobin. *See* Haemoglobin F (HbF)
- Fetal heart rate (FHR), intrapartum monitoring 445
- α -Fetoprotein. *See* AFP (α -fetoprotein)
- Fetoscopy 444*t*
- Fetus
- DNA, analysis 860–861. *See also* Prenatal diagnosis
 - intrapartum monitoring 445
 - maternal drug abuse 485
 - screening for malformation 443–444
 - biochemical screening 443–444, 443*f*
 - current practice 443–444
 - ultrasound 443
 - sex development 412–414, 413*f*
 - thyroid function 379
 - tissue sampling techniques 444, 444*f*, 444*t*
 - well-being, monitoring 445
- Fibre, dietary 192
- cardiovascular disease risk and 757
 - definition 192
 - high intakes 192
 - intake in diabetes 306
 - protective effect against cancer 192, 208
- Fibric acid derivatives (fibrates) 720, 734, 735*t*
- management of diabetes-related cardiovascular risk 308
- Fibrin 408, 507, 508*f*
- Fibrin degradation products (FDPs), measurement 508–509
- Fibrinogen 408, 507, 508*f*
- cardiovascular disease association 753
 - measurement 508, 509*t*
 - properties 572*t*
 - stress response 408
- Fibrinolysis 507
- Fibroblast growth factor 9 (FGF9) 412–413
- Fibroblast growth factor 21 (FGF21), in respiratory chain dysfunction 655
- Fibroblast growth factor 23 (FGF23) 97, 110–111, 112–113, 143
- in chronic kidney disease–mineral and bone disorder (CKD–MBD) 624–625, 626
 - in hyperphosphataemia 113–114
 - in hypophosphataemia 116
 - in hypophosphataemic osteomalacia 623
 - plasma measurements 626
 - in tumour-induced osteomalacia 816
- Fibrocalfic pancreatic diabetes 293
- Fibrofatty lesion, atherosclerosis 743, 743*f*
- Fibrogenesis imperfecta ossium 634
- Fibronectin 447, 607, 607*t*
- Fibrosis
- in carcinoid syndrome 808–809
 - hepatic 244
 - scoring systems for assessment of fibrinogenesis 244, 244*t*
 - serum tests 244–245
 - idiopathic pituitary 367
- FiO₂ (fraction of inspired gas comprising oxygen) 87
- Fish consumption 756
- Fish eye disease 716, 722*t*, 729
- Fish oils 734, 735*t*, 756
- Flat bones 604
- Flavin adenine dinucleotide (FAD) 187, 197–198
- Flavin mononucleotide (FMN) 187
- Flecainide, therapeutic drug monitoring 778
- Fliers 10, 16
- Flora, commensal 218, 561
- Flow cytometry 499–500, 500*f*
- haemolysis investigation 531
- Fluid balance
- in acute kidney injury 140
 - charts 36, 36*t*
- Fluid replacement therapy
- acute kidney injury, prerenal 138
 - diabetic ketoacidosis 328
 - hypernatraemia 45
 - hyperosmolar hyperglycaemic state 329–330
 - sodium deficiency 36, 36*t*
- Fluid requirements, infants *vs.* adults 487*t*
- Fluid restriction, hyponatraemia management 51
- Fluid resuscitation, shocked patients 410
- Flumazenil 794*t*, 803
- Fluoride, dietary 208
- Fluoxetine
- overdose 804
 - therapeutic drug monitoring 780
- Flushing, in carcinoid syndrome 226
- Fluvoxamine
- overdose 804
 - therapeutic drug monitoring 780
- Flux measurements, fatty acid metabolism 657
- Focal neuropathies 691
- Focal segmental glomerulosclerosis 158, 159
- Folate 521, 522*f*
- absorption 521
 - dietary sources 519*t*, 521
 - requirements 521, 522*f*
 - roles 521, 523*t*
 - status, laboratory determination 521
- Folate deficiency 501, 689, 694
- causes 521
 - features 521
 - in malignant disease 816, 816*t*
- Follicle stimulating hormone (FSH) 436
- fluctuations during menstrual cycle 435–436, 435*f*
 - measurement in infertile women 440–441
 - in men with cirrhosis 262
 - secretion 436
 - assessment 352–353
 - effect of oestrogens 439
 - spermatogenesis role 262
- Follicular lymphoma 506–507, 507*f*
- Food
- additives 189
 - contaminants 190
 - enrichment 209–210
 - intake
 - assessment 193
 - effect on test results 8*t*, 9. *See also* Diet(s). *entries beginning dietary*
- Foot ulcers, diabetic 324
- Forearm exercise test 654, 659
- Forensic biochemistry 874–882
- assay request form 875–876, 876*b*
 - chain of custody documents 876, 876*b*
 - commonly requested investigations 875*t*
 - definition 874
 - poisoning with endogenous agents 874, 876–878
 - γ -hydroxybutyrate (GHB) 876–877, 876*b*, 877*b*
 - insulin 877, 877*b*
 - magnesium 875*t*, 877–878, 877*b*
 - sodium 878, 878*b*
 - practical problems 874
 - samples and sampling 875–876. *See also* Post-mortem biochemistry
- 10-Formyl tetrahydrofolate 521, 522*f*
- Forssman antibodies 500
- 14-3-3 Proteins 665, 670–671
- Fractional excretion of sodium (FE_{Na}) 878, 878*b*
- Fractional excretion of water (FE_{H₂O}) 878, 878*b*
- Fractional urinary magnesium (FE_{Mg}), estimation 119
- Fraction of inspired gas comprising oxygen (FiO₂) 87
- Fractures
- alkaline phosphatase activity and 610
 - osteoporotic. *See* Osteoporotic fractures in Paget disease 630
- Fragile-X-associated tremor/ataxia syndrome 700
- Fragile-X syndrome, gene mapping 845*f*
- Frameshift mutation 846–847
- definition 872
- Framingham Heart Study 751
- Frasier syndrome 425
- Frataxin 699
- Fredrickson classification, hyperlipidaemias 721–722
- Free fatty acids
- daily production and elimination 68*t*
 - in diabetic tissue damage 297
 - insulin resistance and 290–291
 - metabolism 68*t*, 69
 - type 2 diabetes and 290–291. *See also* Fatty acid(s)
- Free hormone measurements
- theoretical considerations 383
 - thyroid hormones. *See* Thyroid hormone(s)
- Fresh frozen plasma (FFP) 513
- Friedreich ataxia 699
- Froin syndrome 663–664
- Frozen shoulder 643
- Fructosamine 301
- post-mortem biochemistry 879*t*
- Fructose 220–221
- absorption 221, 221*f*
 - intolerance 270
 - hereditary 270, 464
- Fructose 1,6-diphosphatase deficiency 78, 270
- Fruit consumption, cardiovascular disease risk and 757
- Fruit juices, potassium content 57, 57*t*
- Full blood count (FBC) 497–499, 498*t*
- haemoglobinopathy investigation 557
 - psychiatric disorder investigation 676*t*
- Full crossmatch 512
- Fulminant hepatic failure (FHF) 253, 267*t*
- encephalopathy and 686
 - hypoglycaemia and 345
- Fulminant hepatitis 251
- Fumarylacetoacetate hydrolase deficiency 495
- Fume fever 790*t*
- Functional assessment, nutritional status 194–195
- Functional renal failure (FRF). *See* Hepatorenal syndrome (HRS)
- Fungal poisoning 806, 806*t*
- G**
- Galactosaemia 269, 270*f*, 462–463, 475, 495
- investigations 470, 471, 475
- Galactose 220–221
- absorption 221, 221*f*
 - metabolism 270*f*
 - inherited disorders 495

- Galactose 1-phosphate uridylyltransferase (GALT) 269, 270f
 deficiency. *See* Galactosaemia
 red cell activity, analysis 475
- α -Galactosidase deficiency. *See* Fabry disease
- Gallstones 228, 229t
- Gametes
 definition 872
 formation of 849–850
- Gametogenesis 849–850
- Gangliosides 705
- Garrod, Archibald 844–845
- Gastroectomy, vertical sleeve 206, 206f
- Gastric acid
 loss of 83–84
 protein digestion 223
 secretion 215, 216
- Gastric banding 206, 206f
- Gastric cancer 216
Helicobacter pylori infection and 215, 216
 tumour markers 831
- Gastric emptying 204
- Gastric hypochlorhydria 216, 219
- Gastric juice, fluid composition 34t
- Gastric stretching 204
- Gastric ulcers 215, 216
- Gastrin 215, 216
- Gastrinoma 216
 in multiple endocrine neoplasia type 1 811
- Gastrin releasing peptide, ectopic secretion 815t
- Gastroduodenal ulcers 216
- Gastrointestinal stromal tumours 831–832
- Gastrointestinal tract 214–230
 calcium fluxes 94, 94f
 commensal organisms 218, 561
 disease
 autoimmune 583, 601t
 diet and 208. *See also specific diseases*
 hormones, tumour marker characteristics 823t
 hyperthyroidism manifestations 388
 hypothyroidism manifestations 394
 innate immune system 561
 neuroendocrine tumours 226–227
 normal microflora 218–219
 phosphate fluxes 110f, 111
 potassium loss from 53–54. *See also entries beginning intestinal*
- Gastrostomy, percutaneous endoscopic 210
- Gaucher disease, enzyme replacement therapy 478
- Gaussian distribution 13, 13f
- GBL (γ -butyrolactone) 876–877
- Gender
 assignment 416
 body fluid distribution in relation to 28t
 dysphoria 416
 effect on test results 8, 8t
- Gender differences
 alkaline phosphatase 493f
 cardiovascular disease risk 751–752
 plasma phosphate concentration 113, 113t
 urinary protein excretion 155
- Gender identity 416
 disorder 416
- Gender role 416
- Gene(s) 844–852
 definition 844–846
 expression 846, 847f
 in families and populations 850–851. *See also* Inheritance
 mapping 845, 845f
 mutation. *See* Gene mutations
 structure 845, 846f
 tracking, mutant genes 858–859, 858f
- Gene mutations 846–849, 848f
 analysis. *See* Mutation analysis
 definition 872
 point mutations. *See* Point mutations
- Generalized anxiety disorder 676–677
- Gene therapy 869–871
 β thalassaemia 554
 in cancer 871
 erythropoietic porphyrias 545
 haemophilia 870
 inherited metabolic diseases 481
 stem cells in 870–871
 vectors 870
- Genetic analysis
 applications. *See* DNA analysis, applications
 techniques 855–860
 detection of mutations. *See* Mutation analysis
 detection of specific DNA sequences 855–857
- Genetic diseases 848–849
 bone 632–634, 633t
 variable expression 851–852. *See also* Inherited metabolic diseases (IMDs). *specific diseases*
- Genetic screening 861
 hereditary non-polyposis colon cancer 831
 multiple endocrine neoplasia type 1 812
- Genitalia, external. *See* External genitalia
- Genome, definition 872
- Genotype frequency, calculation 851
- Genotyping
 lipid disorders 734. *See also* Genetic analysis
- Gentamicin, therapeutic drug monitoring 781
- Geophagia 54
- Germ cell, definition 872
- Germ cell tumours 832–834
 cerebrospinal fluid investigations 670
 in children 838
 diagnosis 832
 long-term surveillance 834
 monitoring 833–834, 833f, 834f
 prognosis 833, 833t
 screening 832
- Gestational diabetes mellitus 291–292, 331–332, 446
 recommendations for screening and diagnosis 292b
- Gestational hypertension 38, 156, 764
- Gestational trophoblastic neoplasia (GTN) 834–835
 diagnosis 442, 835
 hyperthyroidism and 394
 monitoring 442, 835
 prognosis 835
 screening 835
- GHB (γ -hydroxybutyrate), forensic biochemistry 876–877, 876b, 877b
- Ghrelin 205
 following bariatric surgery 206
- Gibbs–Donnan effect 27–28
- Gigantism 363
- Gilbert syndrome 245, 246, 247, 494
- Gitelman syndrome 55
 hypomagnesaemia 120t
 non-respiratory alkalosis 84
- Glasgow Coma Scale (GCS) 684, 684t
- Glason score 841
- Gliadin, antibodies to 583. *See also* Coeliac disease
- Glial fibrillary acid protein (GFAP) 665
- Glucagon-related pancreatic polypeptide (GRPP) 205
- Globin chains, haemoglobin 550
 genetic control of synthesis 551–552, 552f
- Glomerular basement membrane (GBM), antibodies to 585, 585t, 586f
- Glomerular filtration 125, 127
- Glomerular filtration rate (GFR) 127
 albumin 153–154, 160
 estimated (eGFR) 134
 definition of CKD using 134–135, 135t
 factors affecting 127b
 measurement 131–136, 132t
 creatinine clearance. *See* Creatinine clearance
 inulin clearance 132, 132t
 isotopic techniques 135–136
 plasma creatinine concentration 132–133, 133f
 in neonates 487, 488
 non-respiratory alkalosis 82
 potassium retention and 59
 in pregnancy 445
 sodium 28, 128. *See also* TmP/GFR
- Glomerular membrane 153, 153f
- Glomerular proteinuria 154, 156–160
 mechanisms 157–158
 pathophysiological consequences 159–160
 tubular proteinuria vs. 160, 160t, 161
- Glomerulonephritis 130, 157
 classification 158–159
 membranoproliferative 158, 159
 urinary sediment 131
- Glomerulosclerosis, focal segmental 158, 159
- Glomerulo-tubular balance 28, 128
- Glomerulus 125, 125f, 126f
 capillary wall 153, 153f
 damage, in diabetic nephropathy 322
 function 127
 development 486–487
 injury 157
- Glucagon
 secretion
 abnormalities in diabetes 295
 ectopic 815t
 hypoglycaemia response 334
 test. *See* Glucagon test
- Glucagon-like peptide-1 (GLP-1) 205, 316
 actions 316, 333–334
- Glucagon-like peptide-1 (GLP-1) analogues 316–317
 adverse effects 317
 hypoglycaemia 317, 341
 mechanism of action 316
- Glucagon-like peptide-2 (GLP-2) 205
- Glucagonoma 227
- Glucagon test 338, 354
 assessment of growth hormone reserve 356
- Glucocerebrosidase deficiency, enzyme replacement therapy 478
- Glucocorticoid(s)
 cytokine interactions 407
 effect on biochemical variables 8t
 effect on calcium/bone metabolism 99, 615
 hyperlipidaemia association 731
 inhibition of thyroid-stimulating hormone release 379, 379f
 resistance, primary 761t
 secretion 352
 effect of radiotherapy 819
 stress response
 cortisol 9, 405, 407
 immune function and 406–409
- Glucocorticoid receptor (GR) 405
- Glucocorticoid replacement therapy following pituitary surgery 359
 Cushing syndrome 366
 monitoring 370

- Glucokinase 277, 277f
loss of function mutations 277
- Gluconeogenesis 68–69, 71, 78f, 274, 333–334
effect of alcohol 321
neonates 488–489, 489f
regulation 274
sites 274. *See also* Glucose metabolism
- Glucose
absorption 221, 221f
blood. *See* Glucose, blood
cerebrospinal fluid. *See* Cerebrospinal fluid
dietary 220–221
emergency hypoglycaemia management 341
homeostasis. *See* Glucose homeostasis
hyperkalaemia management 61
infusion, diabetic ketoacidosis management 328
intolerance. *See* Glucose intolerance
intracellular, fate of 277–278, 277f
measurements
in diabetes mellitus 282, 283b, 283t, 299
post-mortem 879t, 880, 881f
metabolism. *See* Glucose metabolism
oxidation 69, 278. *See also* Glucose metabolism
plasma concentration. *See* Glucose, plasma concentrations
renal tubular handling 168, 169t
resistance 276
storage form. *See* Glycogen
tolerance. *See* Glucose tolerance
urine
measurement 299. *See also* Glycosuria
- Glucose, blood
concentration
high. *See* Hyperglycaemia
low. *See* Hypoglycaemia
normal 273. *See also* Glucose, plasma concentrations
measurement 283t, 299
demonstration of hypoglycaemia 336
factors that may interfere with 300t
in inherited metabolic disease 469–470
in poisoning 875t
in pregnancy 446
psychiatric disorder investigation 676t
in shocked patients 410
sources 274
- Glucose, plasma concentrations
analytical goals 11t
analytical variation 11t
biological variation 11t
critical difference 15t
determinants 274
effect of food intake 9
in hypoglycaemia 334
measurement
diabetes mellitus diagnosis 282, 283b, 283t, 299
factors that may interfere with 300t
in neonates 488–489
in pregnancy 446
- Glucose 6-phosphatase 277, 277f
Glucose 6-phosphatase deficiency 78, 277
Glucose 6-phosphate dehydrogenase 277f, 278, 526–528
- Glucose 6-phosphate dehydrogenase deficiency 278, 494, 526
drugs/chemicals associated with significant haemolysis 526, 527t
gene mapping 845f
heterogeneity 851
tests 531
- Glucose-dependent insulinotropic polypeptide (GIP) 316, 333–334
- Glucose–galactose malabsorption 222
- Glucose homeostasis 273–282
effects of oral contraceptives 448–449
in fed and postabsorptive states 333–334, 334f. *See also* Glucose metabolism; Insulin
- Glucose intolerance
in cirrhosis 263, 295
gestational 446
in malignancy 818
rare conditions associated with 294, 295t. *See also* Diabetes mellitus
- Glucose-lowering therapy, diabetes 309–318
alpha-glucosidase inhibitors 317
bariatric surgery 291, 317–318
DDP-4 inhibitors. *See* Dipeptidyl peptidase IV (DDP-4) inhibitors
immunotherapy 318
insulin. *See* Insulin therapy
intensive glycaemic control 319
islet cell transplantation 318
meglitinides. *See* Meglitinides
metformin. *See* Metformin
pancreatic transplantation 318
peroxisome proliferator-activated receptor γ analogues. *See* Thiazolidinediones
sulfonylureas. *See* Sulfonylureas; Diabetes mellitus, management
- Glucose metabolism
brain 274–275, 334, 662, 687
complete 69
disorders
encephalopathy and 687. *See also specific disorders*
hepatic role 234
incomplete 68–69
intracellular metabolic pathways 277–278, 277f
normal 274–278. *See also* Gluconeogenesis
- Glucose tolerance
impaired. *See* Impaired glucose tolerance (IGT)
in pregnancy 446
- Glucose tolerance test (GTT)
acromegaly diagnosis 362, 372
intravenous 303
oral. *See* Oral glucose tolerance test (OGTT)
- Glucose transporters (GLUTs) 275–276, 276t
assessment of function 303
GLUT1 276, 276t, 289
deficiency 478, 478f
GLUT2 170, 275–276, 276t
GLUT3 275–276, 276t
GLUT4 276, 276t, 333
GLUT5 275–276, 276t
- α -Glucosidase 654–655
inhibitors 317
- Glucotoxicity 291
animal models 278
'Glue sniffing' 788, 803
- Glutamate, metabolism 70, 71, 72f
Glutamate dehydrogenase 71, 490
Glutamic acid decarboxylase (GAD)
antibodies 284, 583
- Glutamine 183
parenteral 211
supplementation 410
synthesis 71
- γ -Glutamyltransferase 241
analytical goals 11t
analytical variation 11t
biological variation 11t
critical difference 15t
liver function tests 241, 248
alcoholic liver disease 241, 257
predictive value, examples 17, 18
in muscle disease 653
- Glutaric aciduria type I 464
- Glutathione metabolism, red blood cells 517, 518f
enzyme defects 526–528, 531
Glutathione peroxidase 191, 199
Glutathione S-transferase (GST) 241
- Glycaemic control
intensive 319
obstacles to achieving 318–322
tests 300–301
- Glycaemic index (GI), cardiovascular disease risk and 756
- Glycated haemoglobin. *See* HbA_{1c} (glycated haemoglobin)
- Glycation, non enzymatic, diabetic neuropathy pathophysiology 692
- Glycine
encephalopathy 670
high plasma concentrations 482
Glycine therapy 482–483
¹⁴C-Glycocholate breath test 219
- Glycogen 274, 334
histocytochemistry 654
Glycogenolysis 234, 277f, 278, 333–334
enzyme investigations 654–655
neonates 488–489, 489f
Glycogen phosphorylase 277f, 278
Glycogen storage diseases 270
investigations, plasma lactate concentrations 474. *See also specific diseases*
- Glycogen synthase 274
- Glycolysis 68–69, 68t, 77, 78f, 274–275
enzyme investigations 654–655
muscle 647, 648f
red blood cells 517, 518f
disorders 528
in tumours 818
- Glycopeptide antibiotics, therapeutic drug monitoring 781
- Glycophosphatidylinositol-anchored HDL-binding protein 1 (GPIHBP1) 721
- Glycosaminoglycans, urinary analysis 474
- Glycosphingolipids 705
- Glycosuria 131, 136, 169–170
hereditary renal 170
- Glycosylated hydroxylysine 611
- Glycosylation 235, 477
congenital disorder of 477
- Glycyrrhetic acid 56
- Glycyrrhizic acid 56
- GM2 gangliosidosis 700
- Goals, analytical 11t, 12
- Goitre
in chronic kidney disease 146
in Graves disease 390
toxic multinodular. *See* Toxic multinodular goitre
- Gonadal dysgenesis, 46XY
complete (Swyer syndrome) 425
partial 425
- Gonadal steroid hormones. *See* Sex hormone(s)
- Gonadotrophin(s)
fluctuations during menstrual cycle 9, 435–436, 435f
in normal puberty 414
pregnancy-related changes 445
secretion 436
assessment 352–353, 356–357. *See also specific gonadotrophins*

- Gonadotrophin deficiency 360, 368–369
 amenorrhoea 353, 368*f*, 369
 associated syndromes 430
 delayed puberty 368–369, 430
 investigations 352–353, 356–357, 368
 females 353, 368, 368*f*
 following pituitary surgery 359
 interpretation of borderline testosterone concentrations 368
 males 352–353
 treatment 368. *See also* Hypogonadism
- Gonadotrophin-dependent precocious puberty 428–429, 428*b*
- Gonadotrophin-releasing hormone (GnRH)
 351–352, 436
 in psychiatric disorders 680
 in puberty 414, 434
 delayed 430
 stimulation test 454, 460
- Gonadotrophin-secreting adenomas 367
- Goodpasture syndrome 585, 585*t*
- Gordon syndrome 60–61, 60*t*, 760*t*, 761*t*
- Gout 639–642
 acute 640
 chronic tophaceous 641
 diabetes association 298
 diagnosis 641, 641*f*
 treatment 641–642
- GPR54 414
- Graft rejection 599, 599*b*
- Graft-versus-host disease (GvHD) 599, 599*b*
- Granular casts, urine 131
- Granulomatous disease
 chronic 579
 hypercalcaemia and 103, 104
- Granulomatous hypophysitis 367
- Graves disease 389, 390*t*
 autoantibodies to thyroidal antigens
 thyroid peroxidase 386, 583
 thyroid stimulating hormone receptor 387, 583
 diagnosis 390, 391*f*
 natural history 391
 ophthalmology 390, 390*f*, 390*t*
 antibodies and 386, 387
 management 391–392
 pregnancy and 380, 387, 391
 skin lesions 390
 thyroid involvement 390
 treatment 391–392
 post-surgery/post-radioiodine
 hypothyroidism 396
- Green tea 757
- Group B streptococcal pneumonia, neonatal 486
- Growth
 delayed, delayed puberty and 369, 430
 in diabetes mellitus 296
 effects of malignant disease treatments 818–819
 increased velocity in normal puberty 416
 retardation, in chronic kidney disease 144–145, 145*b*
 role of vitamin A 185
- Growth factors 573*t*
- Growth hormone (GH)
 in acromegaly 362
 in chronic kidney disease 144–145
 effect of stress 9
 effect on calcium/bone metabolism 99
 effect on insulin 435
 primary insensitivity (Laron syndrome) 368
 in puberty 434
 delayed 430
 reserve, assessment 355–356, 359
- secretion 351
 abnormalities in psychiatric disease 681
 assessment 356
 ectopic 815*t*
 effects of malignant disease therapies 818–819
 hypoglycaemia response 335
 stress response 405
- Growth hormone (GH) deficiency 367–368
 adult 281–282, 360, 368
 effect on insulin sensitivity 281–282
 growth hormone therapy. *See* Growth hormone (GH) therapy
 hypoglycaemia due to 345
 investigations 354, 355–356, 357, 359, 360
 prevention/reversal of microvascular complications in diabetes 297
 re-evaluation of growth hormone status in young adults 356
 weight gain 203
- Growth hormone releasing hormone (GHRH) 351
 in acromegaly 362
 analogues, in malignant disease 819
 in assessment of growth hormone reserve 356
 ectopic secretion 815*t*
 tests 356
- Growth hormone (GH) therapy
 adults, biochemical responses 620
 growth hormone deficiency 367–368
 malignant disease-related 819
 re-evaluation in young adults 356
 non-islet cell tumour hypoglycaemia 344
- Guillain-Barré syndrome 671–672, 692
 antibodies associated with 692*t*
- Gunther disease. *See* Congenital erythropoietic porphyria (CEP)
- Gut. *See* Gastrointestinal tract
- Gut-associated lymphoid tissue (GALT) 562
- Gynaecomastia 457–458
 causes 457, 457*b*, 458*t*
 in dialysed males 146
 investigation 458, 458*f*, 458*t*
- Gyromitrin 806
- H**
- Haem 517, 550
 increased gut content 548
 lyophilized 540
 pigments in cerebrospinal fluid 661, 666–667
 detection 667
 synthesis 533–534, 535*f*
 abbreviations 535*b*
 inherited disorders. *See* Porphyrin(s)
 therapy, acute porphyrias 540–541
- Haem arginate 540–541
- Haematonic studies 500
- Haematocrit
 ECF volume reduction calculation 35, 62
 sodium deficiency 35
- Haematological malignancies 504–507
 hypercalcaemia and 102–103. *See also specific haematological malignancies*
- Haematology 497–501
 blood cell morphology. *See* Blood cell morphology
 definition 497
- Haemochromatosis 265–266, 519
 alcoholism and 257
 articular manifestations 643
 as cause of diabetes mellitus 293
 definition 265
 gene mutations 265
- heterogeneity 851
 investigations 265–266
 transferrin 242, 265–266
 juvenile 265
 male hypogonadism and 262, 456
 neonatal 495
 treatment 266, 643
- Haemodialysis
 acute kidney injury 141
 chronic kidney disease 148–149
 poisoning management 795
 principles 149, 150*f*
 thyroid hormone concentrations and 146
- Haemodilution, in fresh-water drowning 880
- Haemofiltration
 chronic kidney disease 150
 inherited metabolic disease 483
- Haemoglobin 163, 498, 517
 abnormal. *See* Haemoglobinopathies
 buffering 67, 67*f*
 concentration
 adult ranges 498*t*
 measurement 498
 electrophoresis, haemoglobinopathy investigation 557, 558*f*
 fetal. *See* Haemoglobin F (HbF)
 free, haemolysis investigation 530
 function 550–551
 oxygen transport 88–89, 88*f*, 517, 550. *See also* Oxygen–haemoglobin dissociation curve
 glycated. *See* HbA_{1c} (glycated haemoglobin)
 mean cell haemoglobin (MCH) 498
 mean cell haemoglobin concentration (MCHC) 498
 metabolism in haemolytic anaemias 525, 526*f*
 in neonates 485
 post-mortem biochemistry 880
 structural variants 554–557. *See also* Sickle cell disease
 structure 550–551
 synthesis
 genetic control 551–552, 552*f*
 inherited disorders. *See* Haemoglobinopathies
- Haemoglobin A₂ 556, 557, 558
- Haemoglobinaemia 525
- Haemoglobin Bart's hydrops fetalis 552–553, 552*t*, 558
- Haemoglobin Boston 556–557
- Haemoglobin Bristol 556–557
- Haemoglobin C (HbC) 556
- Haemoglobin Chesapeake 556–557, 557*t*
- Haemoglobin D^{Punjab} 556
- Haemoglobin E (HbE) 556
- Haemoglobin F (HbF) 485, 550
 hereditary persistence of 554
- Haemoglobin H (HbH) 558
 inclusions 529
- Haemoglobin H (HbH) disease 552*t*, 553
- Haemoglobin Kansas 556–557, 557*t*
- Haemoglobin Köln 556–557, 557*t*
- Haemoglobin Lepore 554, 557
- Haemoglobinopathies 550–559, 851
 laboratory diagnosis 557–559, 557*t*, 558*f*
 screening 500, 530
 antenatal/preconceptional 558–559
 sickle cell disease. *See* Sickle cell disease
 thalassaemias. *See* Thalassaemia(s)
- Haemoglobin S (HbS) 555, 557*t*
- Haemoglobinuria 162–163, 163*t*, 525, 530
 march 529
- Haemolysis 524–525
 classification of disorders 527*b*

- definition 524–525
diagnosis 529–531
 clinical evidence 529
 laboratory investigations 529–531
 extravascular 524–525
 intravascular 163, 524–525, 529
 laboratory features 525
- Haemolysis, elevated liver enzymes and low platelets (HELLP) syndrome 263, 445, 465–466, 764
- Haemolytic anaemia 516–517
acquired 527*b*, 528–529
 immune haemolytic anaemia 527*b*, 528.
 See also Autoimmune haemolytic anaemia
 non-immune haemolytic anaemia 527*b*, 528–529
autoimmune. *See* Autoimmune haemolytic anaemia
causes 525
chronic 528
haemoglobin metabolism 525, 526*f*
inherited 525–528, 527*b*
 enzyme defects 526–528, 527*b*
 haemoglobinopathies. *See*
 Haemoglobinopathies
 membrane defects 525–526, 527*b*
 in malignant disease 816
 microangiopathic. *See* Microangiopathic haemolytic anaemia
 pathological features 525
 traumatic 528–529. *See also* Haemolysis
- Haemolytic disease of the newborn (HDN) 512–513
- Haemolytic uraemic syndrome 501–503, 503*f*, 529
- Haemopexin 525, 526*f*, 530
- Haemophilia 509
 gene therapy 870
- Haemopoiesis 515–516, 516*f*
- Haemoproteins 533
- Haemorrhage
 intracranial 661–662, 670
 thirst following 32
- Haemorrhagic shock 409
- Haemosiderinuria 525, 530
- Haemosiderosis 265
- Haemostasis 507–510. *See also* entries
 beginning coagulation
- Hair
 excessive growth in females. *See* Hirsutism
 hyperthyroidism manifestations 389
 hypothyroidism manifestations 395
- Hair follicles, effect of dihydrotestosterone 441
- Hallucinations 674
- Haploid cell, definition 872
- Haplotype, definition 872
- Haptocorrins 217
- Haptoglobin 525, 529–530
 properties 572*t*
- Harderoporphyria 542
- Harris–Benedict equation 196, 196*f*
- Hartmann's solution 36*t*
- Hartnup disorder 171–172, 172*b*, 188
- Hashimoto disease 386, 399
 diagnosis 399, 583
- Hashimoto encephalopathy 688
- Haversian canal system 604–605
- HbA_{1c} (glycated haemoglobin)
 cardiovascular disease risk and 754
 diabetic complications risk and 296
 measurement
 diabetes mellitus diagnosis 282, 283, 284*b*
 factors affecting 301
 post-mortem biochemistry 879*t*, 880
 testing of recent glycaemic control 300–301
- hCG. *See* Human chorionic gonadotrophin (hCG)
- Health-associated reference interval 14
- Heart failure 749–750
 congestive 37, 749
 natriuretic peptides as biomarkers. *See*
 Natriuretic peptides (NPs)
 prevalence 749
 in thyroid crisis 388
- Heart rate, energy expenditure and 195–196
- Heart-type fatty acid-binding protein 745*f*, 748
- Heat stroke 882
- Heavy chain diseases 595–596
- Heavy metals
 poisoning 798–799
 antidotes 794*t*
 tubular damage 161
- Height, assessment of nutritional status 193–194, 193*t*
- Heinz bodies 529
- Helicobacter pylori* infection 214
 diagnosis 216
- HELLP ('haemolysis, elevated liver enzymes and low platelets') syndrome 263, 445, 465–466, 764
- Henoch–Schönlein purpura 158
- Heparan sulphate proteoglycans (HSPRG) 713
- Hepatic artery 232, 232*f*
- Hepatic encephalopathy 259, 686, 686*t*
- Hepatic extraction ratio 243–244
- Hepatic (hepatocellular) jaundice 246
- Hepatic lipase (HL) 716
 deficiency 722*t*, 729
- Hepatic myelopathy 689
- Hepatic porphyrias 268. *See also*
 Porphyria(s)
- Hepatic regeneration 234
- Hepatic secretory proteins, measurement 194–195
- Hepatitis flare 255
- Hepatitis
 acute. *See* Hepatitis, acute
 alcoholic 251, 254*t*, 257, 800
 autoimmune. *See* Autoimmune hepatitis (AIH)
 chronic. *See* Hepatitis, chronic
 fulminant 251
- Hepatitis, acute 250–253
 differential diagnosis 251
 viral 252–253, 252*t*
 differential diagnosis 251
 outcome 252–253
- Hepatitis, chronic 253–256
 active 253
 differential diagnosis 254–256, 254*t*
 persistent 253
- Hepatitis A 252, 252*t*
- Hepatitis B virus (HBV) infection 252*t*
 diagnosis 254–255, 254*t*, 255*f*
- Hepatitis C virus (HCV) infection 252*t*
 diagnosis 254*t*, 255
- Hepatitis D virus (HDV) infection 252*t*
- Hepatitis E virus (HEV) infection 252*t*
- Hepatoblastoma 838
- Hepatocellular carcinoma (HCC) 264, 818
 acute hepatic porphyria-associated risk 539
 in children 838
 diagnosis 836
 epidemiology 835
 monitoring 836
 prognosis 836
 screening of high-risk groups 836
 tumour markers 264, 835–836, 838
- Hepatocellular disease
 neonates 495. *See also* Liver disease
- Hepatocyte nuclear factor 1a (HNF-1A), gene mutations 293
- Hepatocytes 232–234, 233*f*
 thyroid hormone entry 377
- Hepatocyte transplantation, inherited metabolic disease management 480
- Hepatoerythropoietic porphyria (HEP) 542–543, 545
- Hepatorenal syndrome (HRS) 139, 261
 differential diagnosis 261, 261*t*
- Hepatotoxicity
 detection 247
 psychotropic drugs 682
- Hepcidin 265, 519, 520*f*, 524
 assays 520
- Hereditary coproporphyrin (HCP) 534*t*, 537*t*, 538
 homozygous 542
 pathophysiology 538
 screening 541*t*. *See also* Porphyria(s), acute
- Hereditary elliptocytosis 526
- Hereditary haemochromatosis. *See*
 Haemochromatosis
- Hereditary non-polyposis colon cancer (HNPCC), genetic pre-screening 831
- Hereditary persistence of fetal haemoglobin (HPFH) 554
- Hereditary renal glycosuria 170
- Hereditary renal hypouricaemia 173–174
- Hereditary spherocytosis 525–526
 diagnosis 531
- Hereditary stomatocytosis (HS) 526
- Hereditary xanthinuria 177–178
- Hereditary ataxia polyneuritis formis 694–695
- Heroin
 abuse 802
 withdrawal, in neonates 485
- Heterophilic antibodies, immunoassay interference 385, 826*b*
- Heterozygote
 compound. *See* Compound heterozygotes
 definition 872
- Hexoaminidase deficiency 700
- Hexokinases 277, 277*f*
- Hexosamine synthetic pathway 277*f*, 278
- Hexosaminidase A deficiency 689
- High bone mass disorders 634
- High density lipoprotein (HDL) 706–707, 708
 assembly 714, 714*f*
 atherosclerosis association 715
 characteristics 708*t*
 metabolism 714–715, 714*f*
 disorders 722*t*, 728–729, 728*b*
 role in reverse cholesterol transport 714*f*, 715
 in nephrotic syndrome 160, 160*t*
 in pregnancy 446
 subclasses 708, 708*t*
 in Tangier disease 696
 in type 2 diabetes 290
- High density lipoprotein (HDL) cholesterol
 cardiovascular disease risk and 752–753
 effect of hormone replacement therapy 449
 effect of menopause 449
 effect of oral contraceptives 448
 inherited disorders affecting 728–729, 728*b*
 measurement 732
- High-dose dexamethasone suppression test 365
- High-dose hook effect 10, 826*b*

- High molecular weight protein markers, renal tubular damage 161–162
- Hip circumference 194
- Hirsutism 441–442
rare causes 442*b*
- Histamine 574–575, 574*t*
- Histochemistry
carbohydrate metabolism disorders 654
respiratory chain defects 655, 655*f*
- HIV infection 579
- HMG-CoA reductase inhibitors. *See* Statin therapy
- Holotranscobalamin 521, 524
- Homeostasis model assessment (HOMA) 303
- Homocystinuria, pyridoxine challenge 479
- Homogentisic acid oxidase deficiency. *See* Alkaptonuria
- Homozygote, definition 872
- Homozygous acute porphyrias 542
- Hook effect
high-dose 10, 826*b*
prolactinomas 362
- Hookworm infestation, iron deficiency anaemia 519–520
- Hormone(s)
appetite control 204–205
cytokines *vs.* 572, 574*t*
ectopic secretion 814, 815*t*
definition 827
gastrointestinal tract, tumour marker characteristics 823*t*
kidney functions involving 144, 144*t*. *See also specific hormones*
- Hormone replacement therapy (HRT)
adrenal disease, monitoring 370–371
pituitary disease, monitoring 370–371
postmenopausal women 447–448, 449–450, 618
adverse effects 449
biochemical responses 619, 619*f*
heart disease and 449–450
hyperlipidaemia association 731
metabolic effects 449
osteoporosis management 450, 618. *See also specific hormones*
- Hormone sensitive lipase (HSL) 713, 717
- Horseshoe kidney 124
- Hot spots 847
definition 872
- Human anti-mouse antibodies (HAMA), assay interference 826*b*
- Human chorionic gonadotrophin (hCG) 436
in ectopic pregnancy 229–230, 442
ectopic secretion 815*t*
in gestational trophoblastic neoplasia 835
hypogonadotrophic hypogonadism treatment 457
in ovarian hyperstimulation syndrome 441
pregnancy diagnosis 442
stimulation test 422, 423*f*, 455, 460
thyroid-stimulating activity 379, 380, 394
as tumour marker 822*t*, 824*t*, 826*t*
germ cell tumours 832, 833, 834
hepatoblastomas 838
- Human epidermal growth factor receptor-2 (HER-2) 823*t*
- Human epididymis protein 4 (HE4) 823*t*
- Human Genome Project (HGP) 845
- Human immunodeficiency virus (HIV) infection 579
- Human leukocyte antigen (HLA) 566, 566*f*, 569–570
antigen presentation 570
class I 569, 569*t*
class II 569–570, 569*t*
genes 569, 569*f*
- Human placental lactogen, ectopic secretion 815*t*
- Humoral hypercalcaemia of malignancy (HMM) 102, 104
- Humoral immunity
primary immunodeficiencies 577–578. *See also* Antibodies; B lymphocytes
- Humoral paraneoplastic syndromes 814
- Hungry bone syndrome 627, 629
- Huntington disease 699
DNA analysis 865–866
- Hurler syndrome (mucopolysaccharidosis type I) 462, 474*t*
- Hyaline membrane disease 485
signs 485–486
- Hyaluronan 606–607
- Hyaluronate 636–637
serum tests for hepatic fibrosis 244
- Hybridization, molecular 852–855
- Hydatiform mole 835
- Hydration, oral, hypernatraemia management 45
- Hydrocephalus 670
- Hydrochloric acid, loss of, non-respiratory alkalosis and 83
- Hydrofluoric acid poisoning, antidote 794*t*
- Hydrogen breath test 219
- Hydrogen ion(s)
activity 65–66
concentration 65, 66
measurement 73
relationship with PCO_2 85, 85*f*
definitions 65–66
excretion. *See* Hydrogen ion excretion
generation 128
in neonates 488
physiological role 65–66
production 68–70, 68*t*
secretion into tubular lumen, mechanisms 169, 169*f*
turnover 68, 68*t*
- Hydrogen ion excretion 70–72, 72*f*
in chronic kidney disease 143
non-respiratory acidosis 75
respiratory acidosis 81
respiratory alkalosis 84
- Hydrogen ion homeostasis 66–73
buffering. *See* Buffering
hydrogen ion excretion. *See* Hydrogen ion excretion
hydrogen ion production 68–70, 68*t*
hydrogen ion turnover 68, 68*t*
in neonates 488
role of liver 72, 73. *See also* Acid–base status
- Hydrogen ion homeostasis, disorders 66, 74–87
acidosis. *See* Acidosis
alcohol-associated 77
alkalosis. *See* Alkalosis
clinical assessment 73
interpretation of acid–base data 85–86, 85*f*
laboratory assessment 73–74
mixed. *See* Mixed acid–base disorders
in pre-term neonates 488
- Hydrogen peroxide, in thyroid hormone synthesis 375*f*, 376
- Hydrostatic pressure 28
- 3-Hydroxy-3-methylglutaryl-CoA reductase inhibitors. *See* Statin therapy
- 3-Hydroxyacyl-CoA dehydrogenases 647–649, 649*f*
measurement of activity 657–658
- Hydroxyapatite 605
- β -Hydroxybutyrate
in ketoacidosis 77, 299, 327
plasma measurements, hypoglycaemia investigation 339
post-mortem biochemistry 879*t*, 880, 881*f*
testing methods 299–300
- γ -Hydroxybutyrate (GHB), forensic biochemistry 876–877, 876*b*, 877*b*
- 25-Hydroxycholecalciferol. *See* 25-Hydroxyvitamin D (25(OH)D)
- Hydroxycobalamin 476–477, 476*f*
- 5-Hydroxyindoleacetic acid (5HIAA) 809, 809*f*
in carcinoid syndrome 809
measurement 809
interfering substances 810*b*
- 17 α -Hydroxylase 437–438
deficiency 353*t*, 370, 425, 760*t*
combined 17,20-lyase deficiency 425, 426*t*. *See also* Congenital adrenal hyperplasia (CAH)
- 1 α -Hydroxylase deficiency 621
- 11 β -Hydroxylase deficiency 353*t*, 423, 424*t*, 760*t*. *See also* Congenital adrenal hyperplasia (CAH)
- 21-Hydroxylase deficiency 353*t*, 370, 419, 423, 424*t*
late-onset 370
sodium wasting 35. *See also* Congenital adrenal hyperplasia (CAH)
- Hydroxyllysine 605
glycosylated 611
- Hydroxylsilylpyridinoline 611–612, 612*f*
- 1-Hydroxymethylbilane (HMB) 533, 535*f*
- 1-Hydroxymethylbilane synthase (HMBS) 533, 535*f*
- 17-Hydroxyprogesterone, measurement, newborns with disorder of sex development 418–419
- Hydroxyproline 611
- 11 β -Hydroxysteroid dehydrogenase (11 β -HSD) 29, 56
- 3 β -Hydroxysteroid dehydrogenase deficiency 353*t*
sodium wasting 35
type 2 424*t*, 425, 426*t*. *See also* Congenital adrenal hyperplasia (CAH)
- 3 β -Hydroxysteroid dehydrogenase isomerase 437
- 17 β -Hydroxysteroid dehydrogenase type 3 (17 β -HSD3) deficiency 425, 426*t*
- 5-Hydroxytryptamine. *See* Serotonin
- 24,25-Hydroxyvitamin D 97*t*
- 25-Hydroxyvitamin D (25(OH)D) 97, 97*t*, 98
plasma concentration
low, osteomalacia and 621, 623
measurement 197, 621–622, 621*t*
seasonal variation 9, 197
- 1 α -Hydroxyvitamin D₃ 98
- 1 α -Hydroxyvitamin D₂ 98
- Hyperacute rejection 599*b*
- Hyperaldosteronism
atrial natriuretic peptide and 38–39
biochemical investigation 762, 765–766
confounding factors 762, 762*b*
interpretation of results 766
patient preparation 765–766
sample preparation 766
screening procedure 766
differential diagnosis 761*t*
familial 55–56
forms of 761–762
glucocorticoid-suppressible 760*t*, 761
hypokalaemic alkalosis 55–56
idiopathic 761
localization of adenoma 762
primary 760–762
- Hyperammonaemia, in inherited metabolic diseases 462, 464, 466, 469, 470
neonatal investigation 470, 471*f*

- Hyperamylasaemia 217–218, 218t
- Hyperbilirubinaemia 237
 in acute hepatitis 251
 in chronic hepatitis 254
 conjugated 246
 further investigation 245
 neonatal/childhood 494–495, 494b
 inherited 245–246
 significance 238–239
 unconjugated 246
 in neonates. *See* Neonates; Bilirubin; Jaundice
- Hypercalcaemia 100–105
 as cause of acute pancreatitis 229t
 causes 101–103, 101b
 in infants and children 104b
 uncommon 103, 103t
 clinical features 101b
 in dialysis-treated patients 625–626, 626b
 differential diagnosis 101b
 encephalopathy and 687
 familial hypocalciuric 102, 102t, 103–104, 121, 491
 investigation 103–104, 105f
 of malignancy 102–103, 104, 814
 in myeloma 594
 neonatal 490–491
 parenteral nutrition-associated 211
 treatment 104–105
- Hypercalciuria 175–176
- Hypercapnia
 encephalopathy and 686–687
 hypoxaemia and 89
 systemic effects 81
- Hypercholesterolaemia 722t
 acquired, causes 730b
 familial. *See* Familial hypercholesterolaemia (FH)
 in hypothyroidism 730
 polygenic 728
 selective screening 4. *See also* Hyperlipidaemia
- Hypercortisolaemia, in depression 679, 680
- Hypercortisolism, articular manifestations 643
- Hyperemesis gravidarum, thyroid function and 380
- Hyperglucagonaemia, in diabetes 295
- Hyperglycaemia
 cerebrospinal glucose concentrations 662
 in chronic liver failure 234
 diabetic neuropathy pathophysiology 691
 hyperosmolar hyperglycaemic state. *See* Hyperosmolar hyperglycaemic state (HHS)
 parenteral nutrition-associated 211
 potassium and 32–33
 prevention 273–274, 275f. *See also* Insulin
 psychotropic drugs causing 682
 rebound 320–321
 stress 273–274, 297–298
 in type 2 diabetes mellitus 287
- Hyperglycaemia-induced hyponatraemia. *See* Hyponatraemia
- Hyperglycinaemia, non-ketotic 482
- Hypericum perforatum*, drug interactions 208–209, 771, 772t
- Hyperinsulinaemia
 neonatal 490
 in type 2 diabetes 287–288
- Hyperinsulinaemia syndrome. *See* Metabolic syndrome
- Hyperinsulinaemic clamps 303
- Hyperinsulinaemic hypoglycaemia, endogenous 343
- Hyperinsulinism, neonatal 489–490
- Hyperkalaemia 32, 57–61
 in acute kidney injury 140
 causes 58–60
 in chronic kidney disease 143
 clinical effects 57–58, 58b
 definition 57–58
 laboratory investigation 60–61
 management 61
 periodic paralysis 59
 redistribution
 in vitro 58, 58b
 in vivo 58–59, 58b
 renal tubular acidosis with. *See* Renal tubular acidosis (RTA), type 4
 secondary renal sodium loss and 35
- Hyperkalaemic periodic paralysis 59, 654
- Hyperlactataemia 69, 78, 91
- Hyperlipidaemia 722t
 acquired (secondary) 729–732
 causes 730b
 drug-related 731–732
 psychotropic drugs 682
 effect of hormonal contraceptives 448
 extravascular manifestations 723t
 familial combined 713–714, 722t, 724, 752
 following renal transplantation 731
 in liver disease 731
 management 734–736
 therapeutic diets 209t, 734
 treatment targets 734
 in nephrotic syndrome 160, 160t, 730
 of pregnancy 446
 WHO classification 721–722, 722t.
See also Hypercholesterolaemia; Hypertriglyceridaemia
- Hyperlipoproteinaemia, remnant 725–726, 726f
- Hypermagnesaemia 121
- Hyponatraemia 44–46
 acute 38, 44
 causes 44, 44b
 definition 39
 encephalopathy and 687
 in hyperosmolar hyperglycaemic state 329, 329b, 330
 management 45–46
 mild 44
 neonatal 487–488
 salt poisoning 878, 878b
 spurious 44
 water deficiency
 with thirst 44
 without thirst 44–46, 45f
 water deficit calculation 62
- Hyperosmolar fluids 36
- Hyperosmolar hyperglycaemic state (HHS) 329–330
 laboratory values, initial 329b
 management 329–330
 presentation and clinical features 329
- Hyperoxaluria 175, 176–177
 in ethylene glycol poisoning 176, 800
- Hyperparathyroidism
 articular manifestations 643
 in chronic kidney disease–mineral and bone disorder (CKD–MBD) 624–625
 classification 96
 clinical presentation 101–102
 familial 101, 102t
 molecular genetics 101, 102t
 neonatal severe 102, 102t
 primary 96, 101–102
 bone disease in. *See* Metabolic bone disease
 in multiple endocrine neoplasia type 1 101, 102t, 810–811
 in multiple endocrine neoplasia type 2 101, 102t
- secondary 96
 treatment 627–628, 629
- Hyperparathyroidism–jaw tumour syndrome 101, 102t, 813
- Hyperphenylalaninaemia 464, 466, 477.
See also Phenylketonuria (PKU)
- Hyperphosphataemia 113–115
 benign transient 240, 247, 493, 610, 851
 isoenzyme analysis 240, 240f
 causes 113–114, 113b
 in chronic kidney disease 142
 consequences 114
 diagnostic approach 114
 in dialysis-treated patients 626, 626b
 therapeutic approach 114–115
- Hyperphosphatasia
 familial benign 610
 familial/idiopathic (juvenile Paget disease) 633t, 634
- Hyperphosphaturia, syndromes of 112
- Hyperproinsulinaemia, familial 279, 280t
- Hyperprolactinaemia
 in anxiety states 681
 in chronic kidney disease
 females 146
 males 145, 146b
 differential diagnosis 360–361, 361f
 in hypothyroidism 395
 management 440
 oligomenorrhoea/amenorrhoea and 361, 440
 presentation 361
 psychotropic drugs causing 681
 secondary diabetes 294
- Hyperpyrexia, malignant 654
- Hyperreninaemia hypoaldosteronism 59, 60–61
- Hypersensitivity reactions 574
 type I 574–575, 580
 investigation 582
 mediators 574, 574t. *See also* Allergies
 type II 575
 type III 575
 type IV 575
- Hypertension 757–765
 in acute porphyria 539
 as cardiovascular disease risk factor 753
 causes 758–759
 secondary. *See* secondary causes (below)
 in chronic kidney disease 142, 759
 definition 757–758
 diabetes in 298
 management 308–309
 dietary salt intake and 756
 haemolytic anaemia associated with 529
 laboratory assessment 759
 malignant 764
 management 764–765. *See also* Antihypertensive therapy
 microalbuminuria and 164
 oral contraceptive use and 449
 portal 259
 in pregnancy 38, 156, 764
 presentation 758
 primary 758–759, 761t
 pulmonary, sickle cell patients 556
 renal hypokalaemic alkalosis with 54b, 55–56
 renovascular 759–760
 secondary causes 758, 758t, 759, 761t
 inherited endocrine disorders 760t
 investigation 759
 pheochromocytoma. *See* Pheochromocytoma
 primary aldosteronism. *See* Hyperaldosteronism
 in sodium excess 36, 38

- Hyperthermia
 in poisoning 793
 post-mortem biochemistry 880–882
- Hyperthyroidism 374–375, 388–394
 anxiety disorders *vs.* 677
 articular manifestations 643
 biochemical parameters increased by 375, 375*b*
 cardiovascular system manifestations 388
 causes 389–394, 390*t*. *See also specific causes*
 central nervous system manifestations 389
 in chronic kidney disease 147
 clinical features 388–389, 388*t*
 in elderly 389
 florid 388
 gastrointestinal system manifestations 388
 hypokalaemic paralysis and 654
 iodine intake and 376
 locomotor system manifestations 389
 neonatal 387, 583
 non-thyroidal illness *vs.* 394
 overt primary, interpretation of thyroid function tests 385
 peripheral nervous system manifestations 389, 693
 renal manifestations 389
 reproductive system manifestations 389
 respiratory system manifestations 389
 secondary 385
 skeletal manifestations 389
 skin and hair manifestations 389
 subclinical 385, 394
 indications for treatment 394, 394*b*
 thyroiditis-associated 398–399
 treatment
 Graves disease 391–392
 in pregnancy 380, 391
 thyroid stimulating hormone-secreting pituitary tumour 393
 toxic adenoma 392
 toxic multinodular goitre 392. *See also Graves disease*
- Hypertonic saline infusion 42, 63–64
 assessment of posterior pituitary function 357
 hyponatraemia
 acute dilutional 51
 chronic dilutional 47, 49*f*, 51
 infusion protocol 63, 63*t*
 pre-infusion preparation 63, 63*t*
- Hypertonic sodium chloride solution 38
- Hypertrichosis, in bullous porphyrias 543
- Hypertriglyceridaemia
 acquired, causes 730*b*
 alcohol intake and 731
 as cause of acute pancreatitis 229*t*
 in chronic kidney disease 730–731
 in diabetes mellitus 729–730
 familial 713–714, 722*t*, 724–725
 lipoprotein lipase gene mutations 716
 management of diabetes-related cardiovascular risk 308. *See also Hyperlipidaemia*
- Hyperlipidaemia 639–642
 asymptomatic 640
 in chronic kidney disease 143
 in malignant disease 816
- Hyperuricaemia 177
- Hyperuricosuria 177
- Hyperventilation
 anxiety-related 677
 non-respiratory acidosis 73, 74–75
 respiratory acidosis 81
- Hyperviscosity 589*t*, 594, 595
- Hypervolaemic hyponatraemia 46, 46*b*, 51
- Hypoadrenalism, assessment
 clinical approach 358–359. *See also Adrenocorticotrophic hormone (ACTH)–adrenal axis, assessment*
- Hypoalbuminaemia
 early onset ataxia with oculomotor apraxia and 700
 in nephrotic syndrome 159
 role in ascites development 259
- Hypoadosteronism
 hyperreninaemic 59, 60–61
 hyporeninaemic. *See Hyporeninaemic hypoadosteronism*
 syndromes of 59–60, 59*t*
- Hypobetalipoproteinaemia 722*t*, 723–724
 familial 722*t*, 724
- Hypocalcaemia 105–109
 causes 106–108, 106*b*
 clinical features 105–106
 differential diagnosis 106*b*
 hypomagnesaemia and 119
 investigation 108–109, 108*f*
 neonatal 108, 491, 491*b*
 parathyroidectomy-associated 108, 627, 629
 treatment 109
- Hypocapnia
 as cause of anxiety 677
 systemic effects 84
- Hypochromic red cells 501, 502*f*, 520
- Hypodipsic hypernatraemia 44–46, 45*f*
- Hypogammaglobulinaemia 579–580
 transient, of infancy 579
- Hypoglycaemia 333–348
 in acute liver failure 234
 alcohol-induced 321, 346–347
 autoimmune 344
 autonomic failure associated with 320
 blood glucose measurements during spontaneous symptoms 336
 causes 341–348. *See also specific causes*
 in cirrhosis 263
 classification 335, 336*b*
 counter-regulatory mechanisms 273, 274*b*
 dawn phenomenon 320–321
 deficient energy intake-associated 347
 definition 334–335
 in diabetes. *See Diabetes mellitus, hypoglycaemia in*
 drug-induced 346
 autoimmune hypoglycaemia 344, 344*b*
 insulin. *See insulin-induced (below)*
 sulfonylureas 315, 340–341, 346
 emergency treatment 341
 encephalopathy and 687
 endocrine deficiency-associated 345
 exercise-induced 321, 340
 factitious/felonious 341
 inherited metabolic disease-associated 348, 469–470
 insulin-induced 340
 alcohol ingestion and 346–347
 correction 320
 investigation 336–341, 337*f*
 demonstration of hypoglycaemia 336–338
 identification of cause 338–339
 in patients with diabetes mellitus 339–341
 in persons without diabetes 336
 liver disease-associated 345
 neonatal 489–490, 490*b*
 neuroendocrine response 320, 334–335
 non-insulinoma pancreatogenous 343
 non-islet cell tumour 282, 345
 as obstacle to achieving glycaemic control in diabetes 319–321
 parenteral nutrition-associated 211
 postprandial 347
- provocation tests 336–338
 reactive 347
 renal impairment-associated 344–345
 response, normal 320, 334–335
 response in diabetic patients 320
 septicaemia-associated 347
 Somogyi effect 320–321
 surreptitious 341
 symptoms 335, 335*b*
 tumour-related 282, 343
 islet cell tumours. *See Insulinomas*
 unawareness 320, 340
- Hypoglycaemia-associated autonomic failure (HAAF) 320
- Hypoglycaemic agents
 as cause of hypoglycaemia 346
 surreptitious administration 341
 screening, hypoglycaemia investigation 339–340. *See also Glucose-lowering therapy, diabetes. specific hypoglycaemic agents*
- Hypogonadism
 hypogonadotrophic. *See Hypogonadotrophic hypogonadism*
 male 455–457
 clinical features 455
 defective hormone synthesis and hormone receptor defects 456
 in hereditary haemochromatosis 262, 456
 primary 455
 secondary 455–456
 treatment 456–457
 primary 431
- Hypogonadotrophic hypogonadism 414, 430, 455–456
 acquired causes 456, 456*b*
 congenital causes 455–456
 idiopathic 456
 treatment 457
- Hypokalaemia 32, 52–57
 causes 52–56
 clinical effects 52, 52*b*
 definition 52
 effect on aldosterone:renin ratio 762
 hypomagnesaemia and 119, 120
 laboratory investigation 56–57
 management 57, 57*t*
 non-respiratory alkalosis 83
 periodic paralysis 52–53
 redistribution
 in vitro 52, 52*b*
 in vivo 52, 52*b*
 in respiratory alkalosis 84
 sodium intake and 38. *See also Potassium depletion*
- Hypokalaemic periodic paralysis 52–53, 654
- Hypokinesia 697
- Hypolipidaemia, acquired 732
- Hypomagnesaemia 118–121
 causes 118
 acquired 118, 119*b*
 genetic 118, 120*t*
 consequences 119
 diabetes association 298
 diagnostic approach 119–121
 hypokalaemia and 119, 120
 therapeutic approach 121, 121*b*
- Hyponatraemia 46–51
 acute. *See Acute dilutional hyponatraemia*
 cerebral salt wasting 50
 chronic. *See Chronic dilutional hyponatraemia*
 classification 46*b*
 definition 39, 46
 diabetes association 298

- encephalopathy and 687
 hyperglycaemia-induced 46
 expected sodium depression estimation 62
 hypervolaemic 46, 46*b*, 51
 hypovolaemic. *See* Hypovolaemic hyponatraemia
 laboratory investigation 50–51
 low osmotic load 49–50
 management 51
 neonatal 487
 psychotropic drugs causing 681
 sick cell syndrome 49
 sodium deficit calculation 62
 spurious 46
 in syndrome of inappropriate antidiuretic hormone secretion (SIADH) 48–49, 49*b*
- Hypoosmolal fluids 36, 36*t*
 Hypoosmolal hyponatraemia 46
 Hypoparathyroidism 106
 autoimmune 106
 genetic causes 106, 107*t*
 neonatal hypocalcaemia 491
 Hypophosphataemia 115–117
 alkaline phosphatase activity and 610
 chronic 694
 consequences 116, 116*t*
 in diabetic ketoacidosis 328
 diagnostic approach 116
 diet and 111
 mechanisms 115–116, 115*t*
 in refeeding syndrome 211–212
 in rickets 492, 623. *See also* Hypophosphataemic rickets
 Hypophosphataemic rickets syndromes of 112
 therapeutic approach 116–117
 Hypophosphataemic osteomalacia. *See* Osteomalacia, phosphopenic
 Hypophosphataemic rickets 492, 622, 622*b*, 623
 autosomal dominant 112, 113
 autosomal recessive 113
 treatment 623
 X-linked 112, 622, 623
 Hypophosphatase 623–624, 624*t*
 Hypophysitis 367
 Hypopituitarism
 diseases leading to 367
 order of progression 354
 thyroid stimulating hormone measurement 352. *See also specific hormone deficiencies*
- Hyporeninaemic hypoaldosteronism 35, 56, 59–61, 59*t*
 in diabetes mellitus 326
 Hypotension
 arginine vasopressin response 31
 hypermagnesaemia and 121
 Hypothalamic disorders
 amenorrhoea 353, 356–357, 368*f*, 369
 polydipsia 41, 46
 Hypothalamic messengers, appetite control 203–204
 Hypothalamo–pituitary–adrenal axis
 abnormalities in psychiatric disease 680
 stress response 404–405
 Hypothalamo–pituitary axis 349
 control of testicular function 452–453, 453*f*
 physiology 350–352, 351*f*
 Hypothalamo–pituitary–gonadal axis 433
 abnormalities in psychiatric disease 680
 evaluation 454–455
 stress response 405
 Hypothalamo–pituitary–thyroid axis,
 abnormalities in psychiatric disease 680
- Hypothalamus
 anatomy 350, 350*f*
 appetite control 203
 deficiency states 367–369
 Hypothermia
 in poisoning 793
 post-mortem biochemistry 880–882
 Hypothyroidism 394–398
 articular manifestations 643
 biochemical parameters decreased by 375*b*
 biochemical parameters increased by 375, 375*b*
 cardiovascular system manifestations 394
 causes 396–397, 396*t*. *See also specific causes*
 central nervous system manifestations 394–395
 in childhood, skeletal effects 395
 in chronic kidney disease 146, 147, 361
 clinical features 394–395, 394*t*
 congenital 394–395, 396–397
 causes 397*b*
 neonatal screening 466, 861
 gastrointestinal system manifestations 394
 in Hashimoto thyroiditis 399
 hyperprolactinaemia and 361
 hypoglycaemia and 345
 iodine and 397
 lipid abnormalities 730
 lithium-induced 397, 681
 locomotor system manifestations 395
 obesity and 203
 overt primary, interpretation of thyroid function tests 385
 peripheral nervous system manifestations 394–395, 693
 postpartum 399
 post-radioiodine 396
 post-surgery 396
 psychiatric manifestations 678, 680
 renal manifestations 395
 reproductive system manifestations 395, 440
 respiratory system manifestations 395
 secondary 385, 397
 as side effect of interferon therapy 255
 skeletal manifestations 395
 skin and hair manifestations 395
 subclinical 385, 398
 treatment 398, 398*b*
 treatment. *See* Thyroid hormone replacement therapy
- Hypouricaemia, hereditary renal 173–174
 Hypoventilation 87
 non-respiratory alkalosis 82
 respiratory alkalosis 84
 Hypovolaemia
 arginine vasopressin response 31
 secondary renal sodium loss and 35
 Hypovolaemic hyponatraemia 46, 46*b*, 51
 sodium deficit calculation 62
 Hypoxaemia 88
 hypercapnia and 89
 in poisoning 788–789
 in respiratory acidosis 81
 Hypoxanthine, post-mortem biochemistry 879*t*
 Hypoxanthine–guanine phosphoribosyl transferase 640*f*
 deficiency 640
 Hypoxia, tissue 90–92
 causes 90*t*
 detection 91
 Hypoxia response elements (HREs) 516
 Hypoxic–ischaemic encephalopathy 687
- I**
 Iatrogenic diabetes 294
 Ibandronate 618
 Idiogenic osmoles 45
 Idiopathic oedema 38
 Idiopathic pituitary fibrosis 367
 Idiotype 568
 Ileal fluid, composition 34*t*
 Ileal segments, urinary diversions 55
 Ileum
 fluid and electrolyte absorption 54–55
 microflora 218
 phosphate absorption 111
 Imaging techniques
 acute kidney injury 138
 aldosterone-producing adenoma 762
 assessment of renal function 130
 carcinoid tumours 810
 central nervous system disorders 668–669
 chronic pancreatitis 218
 gastrinoma localization 216
 insulinoma localization 342
 phaeochromocytoma localization 764
 pituitary gland 360
 thyroid gland. *See* Thyroid gland. *See also specific imaging techniques*
- Imipramine, therapeutic drug monitoring 780
 Immediate hypersensitivity reactions 574–575
 Immediate spin crossmatch 512
 Immobilization, as cause of hypercalcaemia 103*t*
 Immune complex reactions 575
 Immune deficiency 575–580
 infection and 576, 577*b*, 577*t*
 investigation 576–577, 578*f*
 functional assays 602–603
 primary 577–579
 secondary 579–580
 causes 579*t*
- Immune response 560–561
 antigens 561
 clonality 561
 effects of stress 406–409
 exaggerated/persistent. *See* Hypersensitivity reactions
 in malnutrition 195
 neonates 575–576, 576*t*
 primary 560, 568, 568*f*
 secondary 560, 568, 568*f*
 Immune system 560–575
 acute phase proteins. *See* Acute phase proteins
 adaptive. *See* Adaptive immune system
 complement. *See* Complement
 cytokines. *See* Cytokines
 development of immunity 575–576, 576*f*
 diseases 575–588
 allergies. *See* Allergies
 autoimmune diseases. *See* Autoimmune diseases
 B cell malignancies. *See* B cell malignancies
 clinical features 601*t*
 immune deficiency. *See* Immune deficiency
 infections. *See* Infection(s)
 investigations 601*t*, 602–603
 sepsis. *See* Sepsis
 innate 561–562, 562*f*
 mechanisms of immunological damage 574
 Immunity, acquired 560. *See also* Adaptive immune system
 Immunization 560
 Immunofixation 590, 591–592, 592*f*

- Immunoglobulin(s) 561, 566–569, 588
 age-related changes 575–576, 576f
 class switching 568
 constant domains 567
 heavy chain gene 567, 567f
 intrathecal synthesis 664
 in intrauterine infections 485, 485t
 liver function tests 243
 monoclonal. *See* Monoclonal proteins
 polyclonal 568, 588
 properties 568t
 structure 566, 567f
 total, quantification 602
- Immunoglobulin A (IgA) 568
 deficiency 577
 paraproteins 590, 591t
 properties 568, 568t
- Immunoglobulin A (IgA) nephropathy 158
- Immunoglobulin D (IgD)
 paraproteins 582t, 589–590, 591
 properties 568t, 569
- Immunoglobulin E (IgE) 569
 concentrations, raised 581
 measurement 602
 allergy investigation 581, 581t, 582t
 properties 568t
 specific 581
- Immunoglobulin E (IgE)-mediated hypersensitivity 574, 580
 mediators 574t. *See also* Allergies
- Immunoglobulin G (IgG) 568, 568f
 antigen-specific, quantification 602
 in autoimmune hepatitis 256
 cerebrospinal fluid 663, 663f, 663t
 oligoclonal banding 664, 665f, 669–670, 671
 clearance 154t
 liver function tests 243
 neonates 575–576, 576f
 paraproteins 591t, 592f
 properties 568, 568t
 subclasses 568
 measurements 602
 urine protein selectivity studies 159
- Immunoglobulin M (IgM) 568, 568f
 in intrauterine infections 485, 485t
 liver function tests 243
 paraproteins 590, 591t, 595
 properties 568t
 X-linked hyper-IgM syndrome 578
- Immunological markers, renal function 130
- Immunometric assays (IMAs), thyroid stimulating hormone 383
- Immunomodulating therapies, shocked patients 411
- Immunophenotyping 562–564
- Immunoreactive trypsinogen (IRT) 217–218
- Immunosuppressive therapy
 hyperlipidaemia association 732
 therapeutic drug monitoring 783–785
 transplant patients 600t
 infection secondary to 599b
 renal transplantation 150
- Impaired fasting glycaemia (IFG), diagnosis 282–283, 283t
- Impaired glucose tolerance (IGT)
 as cardiovascular disease risk factor 754, 754f
 in chronic kidney disease 148, 148b
 diagnosis 282–283, 283t
- Impotence 458–459
 causes 458, 458b, 459b
 in chronic kidney disease 145, 145b
 investigation 458–459, 459b
 treatment 459
- Imprecision, analytical 9–10, 11, 11t
- Imprinting 851–852
 definition 872
- Inborn errors of metabolism. *See* Inherited metabolic diseases (IMDs)
- Incretin effect 279, 316
- Incretin mimetics 316
- Indian childhood cirrhosis 267t, 268
- Indirect antiglobulin test (IAT) 511
- ¹¹¹Indium white cell technique 225
- Individual screening 4–5, 861
- Indocyanine green, clearance tests 244
- Inducible degrader of low density lipoprotein receptor (IDOL) 720
- Infancy, transient hypogammaglobulinaemia of 579
- Infantile hypercalcaemia (Williams syndrome) 491
- Infants
 aldosterone reference ranges 766
 hypercalcaemia, causes 104b
 inherited metabolic disease presentation 464
 maternal diabetes-associated complications 331b
 mortality rates 484
 poisoning 788
 sudden infant death syndrome 882, 882t
 thyroid function 379
 vitamin K deficiency bleeding 186–187.
See also Neonates
- Infection(s) 597–598
 in acute kidney injury 140
 cardiovascular disease risk and 755
 as cause of acute pancreatitis 229t
 central nervous system
 acute 669
 chronic 669–670
 diagnosis and monitoring 598
 haemolytic anaemia-associated 527b, 528
 immune deficiency and 576, 577b, 577t
 impaired resistance in chronic energy deficiency 201
 intrauterine 484–485, 485t
 recurrent, investigations 601t
 renal stone formation 177
 secondary to immune suppression 599b.
See also Bacterial infections
- Infectious mononucleosis 500–501, 505f
- Inferior petrosal sinus sampling (IPSS), adrenocorticotrophic hormone 365–366
- Infertility
 definition 440
 female. *See* Female infertility
 male. *See* Male infertility
- Inflammation
 cardiovascular disease risk and 755
 cerebrospinal fluid markers 667–668
 microalbuminuria as risk factor 165
 post-mortem biochemistry 882
 stages 571–572
 stimulators 571–572
 stress and 406–409
- Inflammatory arthritis 637–638. *See also* Rheumatoid arthritis (RA)
- Inflammatory bowel disease (IBD) 215
 bone mineral density 220
 faecal calprotectin concentrations 225–226
- Inflammatory cytokines 566, 574
- Inflammatory diseases
 as cause of hypopituitarism 367
 central nervous system 701. *See also specific diseases*
- Inflammatory mediators 571–572, 572t
- Information systems 23
- Inheritance 466–468, 850–851, 850f
 autosomal dominant 467, 467f, 850, 850f
 autosomal recessive 466–467, 467f, 850, 850f
 mitochondrial 467–468, 468f, 850
 X-linked 467, 468f, 850, 850f
- Inherited metabolic diseases (IMDs)
 461–483
 clinical presentation 461–466
 during adulthood 465
 in later infancy 464
 neonatal. *See* Neonates
 postpartum 466
 during pregnancy 465–466
 at puberty 464–465
 at weaning 464
 diagnostic strategies 468–476
 cerebrospinal fluid examination 670
 confirmatory investigations 475–476
 essential laboratory investigations 469–471
 functional and loading tests 474–475
 second-line investigations 472–474
- DNA analysis 862–866
 diagnosis of index cases 860
 enzyme analysis 475
 hepatic 265–270
 hypoglycaemia association 348
 inheritance. *See* Inheritance management 476–483
 blockage of site of action of toxic metabolites 482
 enzyme replacement therapy 478–479
 gene therapy 481
 inhibition of product breakdown 478
 molecular therapies 481
 nutritional support 476–478
 organ transplantation 479–480
 product replacement strategies 476–478
 substrate depletion 483
 substrate deprivation 483
 toxic metabolite reduction strategies 481–482
 toxic substance removal strategies 482–483
 newborn screening 466. *See also* Neonatal screening
 pathophysiology 461–466
 prenatal diagnosis 476. *See also specific diseases*
- Inhibin(s) 436–437
 measurement, in disorder of sex development (DSD) 421–422
 as tumour markers 823t, 838, 841–842
- Inhibin A 421–422, 436
 tumour marker characteristics 823t
- Inhibin B 414, 421–422, 436
 tumour marker characteristics 823t
- Innate immune system 561–562, 562f
- Inositol triphosphate 93
- Insertional mutagenesis 870
- Insulin 234, 278–281
 actions 279–281, 280t, 333
 antibodies 295
 in autoimmune hypoglycaemia 344, 344b
 hypoglycaemia investigation 339
 antihyperglycaemic actions 275f
 appetite control 204
 biosynthesis 278–279
 abnormalities 279, 280t
 deficiency
 diabetic ketoacidosis development 326, 327f
 in type 1 diabetes mellitus 283

- in type 2 diabetes mellitus 286–288. *See also* Diabetes mellitus
 - effect of food intake 9
 - effect on potassium distribution 32
 - exogenous administration, forensic
 - biochemistry 877, 877b
 - GLUT4 responsiveness 276
 - hyperkalaemia management 61
 - internalization 281
 - in liver failure 234
 - in obese patients 8
 - pharmacokinetics 279
 - plasma concentrations 279
 - hypoglycaemia investigation 337f, 338–339
 - receptor. *See* Insulin receptor
 - requirements in pregnancy 331, 446
 - resistance. *See* Insulin resistance
 - second messenger systems 281
 - secretion 279
 - abnormalities 279, 285, 286–287
 - effects of menopause 449
 - postprandial 333–334
 - in protein–energy malnutrition 201
 - therapy. *See* Insulin therapy
 - tolerance test. *See* Insulin stress test (IST)
- Insulin-binding antibodies 295, 301
- Insulin-degrading enzyme (IDE) 877
- Insulin-like factor 3 (INSL3), measurement
 - in disorder of sex development 420
- Insulin-like growth factor(s) (IGFs) 279–280, 281–282
 - in non-islet cell tumour hypoglycaemia 343–344
 - physiological role 281
- Insulin-like growth factor 1 (IGF-1) 281, 351
 - in acromegaly 362
 - in diabetic retinopathy 282
 - effect on TmP/GFR 111–112
 - in growth hormone deficiency 354
 - assessment 356
 - in sex steroid regulation 435
- Insulin-like growth factor 2 (IGF-2) 281
 - ectopic secretion 815t
 - in tumour-related hypoglycaemia 282, 343–344
- Insulin-like growth factor (IGF)-binding proteins, in non-islet cell tumour hypoglycaemia 343–344
- Insulinomas 227, 341–343
 - clinical features 342
 - diagnosis 342
 - localization 342
 - treatment 342–343
- Insulin receptor 279–280, 281, 281f
 - antibodies, in autoimmune hypoglycaemia 344, 344b
 - gene mutations 281
- Insulin resistance 278, 679
 - in chronic kidney disease 148
 - in cirrhosis 263
 - definition 288
 - in diabetes mellitus
 - type 1 285–286
 - type 2 286–287, 288
 - ghrelin concentrations and 205
 - in malignancy 818
 - in obesity 289
 - severe 294–295
 - tests 301–303. *See also* Metabolic syndrome
- Insulin stress test (IST) 354, 355, 371–372
 - assessment of growth hormone reserve 355–356
 - contraindications 354, 355–356
- Insulin therapy 310–313
 - administration 311–313
 - continuous subcutaneous insulin infusion 311, 312t
 - diabetic ketoacidosis management 328
 - effect of exercise 306
 - effect on blood glucose concentration, factors affecting 319t
 - gestational diabetes 292
 - hospital setting 330
 - hyperosmolar hyperglycaemic state 329–330
 - hypoglycaemia associated with. *See* Hypoglycaemia, insulin-induced
 - insulin analogues 310
 - intermediate-acting insulin 310
 - long-acting insulin analogues 311
 - premixed insulin analogues 310–311
 - regimens 311, 312t
 - regular insulin 310
 - type 1 diabetes 310–313
 - intensive 319
 - type 2 diabetes 313, 317
 - variable infusion rates (sliding scales) 328
- Insulin tolerance test (ITT). *See* Insulin stress test (IST)
- Interactions, drug. *See* Drug interactions
- Intercalated cells 126
- Interface hepatitis 254
- Interference 11
- Interferons 573t
- Interferon therapy, hypothyroidism as side effect 255
- Interindividual variation 10, 11t
- Interleukin (IL) 573t
- Interleukin-1 (IL-1) 573t
 - stress response 406
- Interleukin-6 (IL-6) 573t
 - stress response 406, 408
- Intermediate-acting insulin 310
- Intermediate density lipoprotein (IDL) 706–707, 708, 708t, 714
- Intermediate syndrome 799
- Internal quality control (IQC) 21
 - tumour marker measurements 826
- International normalized ratio (INR) 242
 - in poisoning 875t
- Intersex 416. *See also* Disorder of sex development (DSD)
- Interstitial nephritis 160–161
 - hyperkalaemia 60
- Interstitial space 27
- Intestinal absorption
 - calcium. *See* Calcium
 - carbohydrates. *See* Carbohydrate(s), absorption
 - cholesterol 710–711
 - disorders. *See* Malabsorption
 - folate 521
 - iron 519
 - magnesium 117, 118, 118f
 - oxalate, increased 176
 - phosphate 110f, 111
 - proteins. *See* Protein(s), absorption
 - triglycerides 224, 711
 - vitamin B₁₂ 521
- Intestinal carcinoid tumours 226
- Intestinal decontamination 794
- Intestinal inflammation, faecal tests 225–227
- Intestinal microflora 218–219, 561
- Intestinal permeability 225
- Intoxication, inherited metabolic disease
 - presentation 462–463
- Intracellular fluid (ICF) 27–28, 28t
 - potassium and 28t, 32–33
 - water and 30–32
- Intracellular signalling, calcium role 93
- Intracranial haemorrhage 661–662, 670
- Intracytoplasmic sperm injection (ICSI) 457
- Intrahepatic cholestasis of pregnancy 262–263, 445
- Intraindividual variation 10, 11t
- Intrapartum fetal monitoring 445
- Intrathecal immunoglobulin synthesis 664
- Intrauterine infections 484–485, 485t
- Intrauterine poisoning 788
- Intravascular haemolysis 163
- Intravascular space 27
- Intravenous glucose tolerance testing (IVGTT) 303
- Intrinsic acute kidney injury. *See* Acute kidney injury (AKI)
- Intrinsic biological variation 9–10, 11t, 15
- Intrinsic factor 215, 217, 521
 - autoantibodies 217, 523, 583
- Intron 845
 - definition 872
- Inulin clearance 132, 132t
- Invasive moles 835
- In vitro fertilization (IVF) 441
 - preimplantation genetic diagnosis 444, 444t
- Iodide, incorporation into thyroglobulin 375, 375f
- Iodine
 - deficiency 376, 397
 - as cause of hypothyroidism 397
 - excess, effects 376
 - hyperthyroidism 390t, 393
 - hypothyroidism 397
 - thyroid hormone synthesis and 376
- Iodine-131 therapy. *See* Radioiodine therapy
- Iodothyronine deiodinases 377, 378f
 - deficiency in congenital hypothyroidism 396–397
- Ion exchange chromatography, in investigation of amino acid disorders 472
- Ion–semiconductor sequencing 860
- Ipecacuanha, syrup of 794
- Iron 517–519
 - absorption 519, 520f
 - deficiency
 - clinical consequences 520. *See also* Iron deficiency anaemia
 - dietary sources 519t
 - excretion 265
 - functions 517–519
 - overload 265–266
 - alcoholic liver disease 257
 - transferrin concentration 242. *See also* Haemochromatosis
 - poisoning. *See* Iron poisoning
 - requirements 519, 519t
 - serum, measurement 520
 - status, laboratory investigation 520
 - transport and storage 519
 - Iron deficiency anaemia 517–519
 - causes 519–520
 - clinical features 520
 - laboratory investigation 520
 - in rheumatoid arthritis 644
 - in malignant disease 816t
 - morphological features 501, 502f
 - Iron poisoning 792t, 798
 - analysis 792t, 798
 - clinical features 798
 - mechanisms 798
 - treatment 794t, 798
 - Iron response elements (IREs) 519
 - Iron response proteins (IRPs) 519
 - Irreversibly sickled cells (ISCs) 555

- Irritable bowel syndrome (IBS) 675
 faecal calprotectin concentrations 226
 prevalence 214
 small bowel bacterial overgrowth 219
 Islet cell antibodies (ICA) 284, 583
 Islet cell transplantation 318
 Islet cell tumours 226–227
 as cause of hypercalcaemia 103*t*. *See also*
 Insulinomas
- Isoelectric focusing (IEF), oligoclonal band
 detection 664, 665*f*; 669–670, 671
- Isoleucine, metabolism 480*f*
- Isoniazid, therapeutic drug monitoring 782
- Iso-osmolal fluids, sodium deficiency
 management 36, 36*t*
- Isotope uptake studies, thyroid gland. *See*
 Thyroid scintiscanning
- Isotopic techniques, glomerular filtration rate
 measurement 135–136
- J**
- Jamaican neuritis 694
- Jaundice 237, 238
 acholuric 245
 in Crigler–Najjar syndrome 246
 differential diagnosis 245
 in Gilbert syndrome 246
 hepatic (hepatocellular) 246
 in inherited metabolic diseases 470
 neonatal (physiological) 247, 493–494
 post-hepatic (cholestatic) 246
 pre-hepatic 245. *See also*
 Hyperbilirubinaemia
- J-chain 568
- Jejunal segments, urinary diversions 55, 59
- Jejunocolic anastomosis 213
- Jejunum
 fluid and electrolyte absorption 54–55
 phosphate absorption 111
- Jod–Basedow phenomenon 376, 393
- Joints 636–637
 cartilaginous 636
 disorders. *See* Articular disorders
 synovial 636–637, 637*f*
- Juvenile haemochromatosis 265
- Juvenile Paget disease 633*t*, 634
- K**
- Kaliuresis 57
- Kallmann syndrome 360, 368, 430, 455
- Kava (*Piper methysticum*) 208–209
- Kayser–Fleischer rings 255, 266, 698–699
- Kearns–Sayre syndrome 655, 656
- Kell antibody 511
- Kenny–Caffey syndrome 107*t*
- Kernicterus 246, 247, 493
- Keshan disease 191
- Ketoacidosis 76–77
 alcoholic 76–77, 330, 800
 diabetic ketoacidosis *vs.* *See* Diabetic
 ketoacidosis
 diabetic. *See* Diabetic ketoacidosis
- Ketoacids, metabolism 68*t*, 69
- Ketogenesis 69
 in alcoholic ketoacidosis 330
- Ketogenic diet 478, 478*f*
- Ketone bodies 327, 657
 effects of fasting 339, 339*f*
 utilization by brain 274–275
- Ketones
 metabolism in neonates 489
 post-mortem biochemistry 880, 881*f*
- Ketone testing 299–300, 322, 327
 in inherited metabolic diseases 469, 470
- Ketonuria 327
 in inherited metabolic diseases 469, 470
- Ketosis 469, 470
- Key performance indicators, laboratory
 medicine 24–25, 24*t*
- Kidd antibodies 511
- Kidney(s)
 anatomy 124–127
 anomalies in children with disorder of sex
 development 417
 bicarbonate excretion, non-respiratory
 alkalosis 82–83
 blood flow. *See* Renal blood flow
 calcium fluxes 94–95, 94*f*. *See also* Calcium,
 analysis of tubular handling
 control of potassium output 33
 control of sodium output 28–30, 128
 control of water output 30–31
 damage, diabetic 298–299
 development 486–487
 diseases affecting 130. *See also* Renal
 disease
 function. *See* Renal function
 gross anatomy 124
 hydrogen ion excretion. *See* Hydrogen ion
 excretion
 hyperthyroidism manifestations 389
 hypothyroidism manifestations 395
 magnesium homeostasis 117–118, 118*f*.
See also Magnesium, renal tubular
 reabsorption
 microstructure 124–127, 125*f*
 parathyroid hormone actions on 96
 phosphate reabsorption 110*f*, 111–113
 protein conservation 152–155
 responsiveness to arginine vasopressin 31,
 32*f*
 stress response 406. *See also* entries beginning
 renal
- Kisspeptin 414
- KIT protein 831–832
- Kleinhauer test 512–513, 513*f*
- Klinefelter syndrome 431, 455
- Klotho 143
- Korsakoff psychosis 187, 686
- K-RAS* mutation analysis
 colorectal cancer 831
 lung cancer 837
- Krebs–Henseleit cycle 235
- Kringles 710
- Kussmaul breathing 74–75
- Kwashiorkor 201
- L**
- Labioscrotal fusion 413–414, 417, 418*f*
- Laboratory error, forensic investigation
 874–875, 875*b*
- Laboratory medicine, quality aspects 21–26
 clinical quality indicators 24–25, 24*t*
 demand management 25
 evaluation/audit of services 23, 23*b*, 23*f*
 information systems 23
 personnel 22
 premises and environment 22
 quality assurance. *See* Quality assurance
 quality management systems 22
 quality standards 21–23
 regulation of laboratories 22, 22*b*
- Labour 446–447
 fetal monitoring 445
 pre-term 447
- Lactase 221, 222
 deficiency 220, 222
- Lactate
 blood measurements
 in inherited metabolic diseases 473–474
 in poisoning 875*t*
 in respiratory chain dysfunction 655
 cerebrospinal fluid. *See* Cerebrospinal fluid
 metabolism 68–69, 68*t*, 77, 278
 normal plasma concentration 77
 post-mortem biochemistry 879*t*, 880
- Lactate dehydrogenase (LDH) 654–655
 in cerebrospinal fluid 667
 tumour marker characteristics 822*t*
- Lactic acid 77
 isomers 78
- Lactic acidosis 77–78
 causes 77, 77*b*
 inherited metabolic disorders 463, 463*b*,
 473, 483
 metformin-associated 78, 313–314
 type A 77*b*, 78
 type B 77*b*, 78
- Lactose 220–221
- Lactose tolerance test 222
- Lambert–Eaton syndrome 814
- Lamellar bone 604–605
- Lamina densa 153, 153*f*
- Lamina rara externa 153, 153*f*
- Lamina rara interna 153, 153*f*
- Lance–Adams syndrome 687
- Landsteiner's law 510
- Laparoscopy, evaluation of internal
 anatomy in disorder of sex
 development 418
- Large intestine. *See* Colon
- Laron syndrome 368
- Lasofloxifene 618
- Latent autoimmune diabetes of adults
 (LADA) 283–284
- Latex agglutination test 500
- Laxative abuse 55
- LDL. *See* Low density lipoprotein (LDL)
- Lead
 poisoning 792*t*, 798–799
 tubular damage 161
- Leber hereditary optic neuroretinopathy
 465, 656
- Lecithin, fetal lung production 444–445
- Lecithin–cholesterol acyltransferase (LCAT)
 708, 709, 715, 716
 in atherosclerosis 716
 deficiency
 familial 716, 722*t*, 729
 partial 716, 722*t*, 729
 in nephrotic syndrome 160
- Lectin pathway, complement activation 570,
 571*f*
- Leprechaunism 281, 295
- Leptin 204–205
 deficiency 202, 204–205
 effect on bone metabolism 99
- Lesch–Nyhan syndrome 640
 gene mapping 845*f*
- Leukaemia 588
 acute 504–505, 505*f*
 myelomonocytic, renal potassium loss
 56
 chronic 505–506
 lymphocytic 506*f*, 591*t*, 595–596
 myeloid 505, 506*f*
- Leukocyte adhesion defects 579
- Leukocytes. *See* White blood cells (WBCs)
- Leukocytosis 503–504
 cerebrospinal fluid 662
- Leukoerythroblastic anaemia 524
- Leukopenia 504
- Leukotrienes 705, 706*f*
- Levothyroxine therapy. *See* Thyroid
 hormone replacement therapy

- Leydig cell(s) 451
 in chronic kidney disease 145–146
 development 412–413, 413f
 hyperplasia 429
 hypoplasia 425, 426t
 tumours 841–842
- Liddle syndrome 56, 760t, 761t
- Lifestyle modifications, osteoporosis management 617
- Li–Fraumeni syndrome 868
- Ligase-mediated allele detection 855–856, 856f
- Light chain proteinuria 163–164, 163t
- Lignin 192
- Likelihood ratios 19–20
- Linkage 849, 849f
 definition 872
- Linkage analysis 858, 858f
 polymorphic markers 858, 859f
- Linkage disequilibrium 849
 definition 872
- Linoleic acid 182–183, 183f
- α -Linolenic acid 182–183, 183f, 756
- Lipase(s) 224, 716–718
 lipoprotein. *See* Lipoprotein lipase
 lysosomal acid 717
 pancreatic 217–218, 224
 pancreatic function testing 217–218
 acute pancreatitis 229. *See also specific lipases*
- Lipase maturation factor 1 717
- Lipid(s) 703–706
 function 703t
 major classes 703t
 nuclear 706
 plasma
 cardiovascular disease risk and 752–753
 pregnancy-related changes 446
 rafts 703–704. *See also specific lipids*
- Lipid disorders 736
 investigation 732–734. *See also Dyslipidaemia. specific disorders*
- Lipid-lowering agents 734
 management of diabetes-related cardiovascular risk 308. *See also Statin therapy*
- Lipid metabolism
 in chronic kidney disease 148, 730–731
 effects of hormonal contraceptives 448
 effects of hormone replacement therapy 449
 effects of menopause 449
 hepatic 234–235
 in malignancy 818
 stress response 405, 407
- Lipid oxidation hypothesis 741–743, 742f
- Lipid preparations, parenteral nutrition 211
- Lipodystrophy 295
- Lipolysis
 in adipose tissue 713
 effects of hypoglycaemia 339, 339f
- Lipoprotein(a) 707t, 708, 710
 measurement, investigation of lipid disorders 733
- Lipoprotein(s) 703, 706–708
 assembly 711–713, 712f, 714, 714f
 classes 706–707, 708t
 disorders
 classification 721–723, 722t. *See also specific disorders*
 high density. *See* High density lipoprotein (HDL)
 intermediate density 706–707, 708, 708t, 714
 low density. *See* Low density lipoprotein (LDL)
- metabolism. *See* Lipoprotein metabolism
 in nephrotic syndrome 160, 160t
 plasma, cardiovascular disease risk and 752–753
 pregnancy-related changes 446
 separation techniques 733–734
 structure 706–707, 707f
 very low density. *See* Very low density lipoprotein (VLDL)
- Lipoprotein-associated phospholipase A2 (LpPLA2) 717–718
- Lipoprotein lipase 713, 716
 deficiency 725
 inhibitor, familial 725
- Lipoprotein metabolism 711–715, 712f
 endogenous pathway 713–714
 enzymes involved in 716–718
 exogenous pathway 713
 hepatic cholesterol trafficking 714
 lipolysis in adipose tissue 713
 receptors involved in 718–720
 transfer proteins involved in 718
- Lipoprotein-related receptor 5 (LRP5) 607
- Lipoprotein X 708, 731
- Liposomes, in gene therapy 870
- Lipotoxicity 290–291
- Liraglutide 316, 317
- Lithium
 metabolic complications 681
 hypercalcaemia 103t
 hypothyroidism 397, 681
 nephrogenic diabetes insipidus and 41, 42
 poisoning 792t, 804–805
 therapeutic drug monitoring 770–771, 780
- Liver
 anatomy 232–234, 232f, 233f
 biotransformation 235
 cancer, primary. *See* Hepatocellular carcinoma (HCC)
 carbohydrate metabolism 234
 cholesterol metabolism 236, 236f, 703, 711
 cholesterol trafficking 714
 circulation 232
 drug-induced damage 254–256, 263t
 effects of parenteral feeding 212
 encephalopathy 259
 excretion 235
 fatty. *See* Steatosis
 function
 psychotropic drugs interfering with 682
 tests. *See* Liver function tests
 glucose output 274
 effect of insulin 279–280. *See also* Gluconeogenesis
 insulin actions 280t
 insulin clearance 279
 lipid metabolism 234–235
 lobes 232, 232f
 macroscopic structure 232, 232f
 microscopic structure 232–233, 233f
 physiological functions 234–236
 protein metabolism 235
 regeneration 234
 role in hydrogen ion homeostasis 72, 73
 secretory proteins, measurement 194–195
 ultrastructure 233–234, 233f. *See also entries beginning hepatic*
- Liver disease 250–272
 alcoholic. *See* Alcoholic liver disease
 autoimmune 584–585, 584t, 601t
 chronic
 assessment of prognosis 248
 carbohydrate metabolism 234
 encephalopathy and 686
 monitoring response to therapy 247
 peripheral neuropathy and 693
 progression from viral hepatitis 252
 sexual dysfunction 261–262
 sodium retention 37. *See also* Cirrhosis
 classification 250, 251b
 drug-induced 254–256, 263t
 hypoglycaemia associated with 345
 inherited metabolic disorders 265–270
 lipid abnormalities 731
 monitoring response to therapy 246–247
 neonatal 493–496
 in inherited metabolic diseases 470, 471b
 neoplastic 264. *See also* Hepatocellular carcinoma (HCC)
 normal liver function tests 248
 in older children 495–496
 parenteral nutrition-induced 212
 peripheral neuropathy and 693
 in pregnancy 263
 protoporphyrin 547
 severity assessment 248, 248t. *See also specific diseases*
- Liver failure
 acute. *See* Acute liver failure (ALF)
 encephalopathy and 259, 686, 686t
- Liver function tests 231, 236–245
 abnormal, in asymptomatic patients 247–248
 in acute hepatitis 250–251
 viral 252
 in acute liver failure 253
 applications 236
 bacterial infection-associated abnormalities 264–265
 bile acids 243
 bilirubin and bile pigment metabolism 237–239, 248
 in chronic hepatitis 254
 clearance tests 243–244
 in hepatic encephalopathy 686
 in inherited metabolic diseases 470
 in neoplastic liver disease 264
 normal, in presence of overt liver disease 248
 parenteral nutrition-associated abnormalities 264
 plasma enzyme activities 239–241
 in alcoholic liver disease 258
 alkaline phosphatase. *See* Alkaline phosphatase (ALP)
 aminotransferases. *See* Aminotransferases
 γ -glutamyltransferase. *See* γ -Glutamyltransferase
 glutathione S-transferase 241
 in pregnancy 262–263
 plasma proteins 241–243
 acute phase reactants 243
 albumin 242
 α 1-antitrypsin 242
 caeruloplasmin 243
 coagulation factors 242
 α -fetoprotein 242
 immunoglobulins 243
 transferrin 242–243
 pregnancy-related changes 262–263
 in primary biliary cirrhosis 256
 quantitative evaluation 243–244
 role in assessing prognosis 248–249
 serum tests for hepatic fibrosis 244–245
 uses 245–247
- Liver–kidney microsomal (LKM) antibodies 255–256

- Liver transplantation 248, 270–272
 acute graft rejection 271
 acute liver failure management 253, 270
 acute porphyria management 541
 immediate postoperative period 271
 indications 270
 inherited metabolic disease management 479
 intermediate follow-up 271
 laboratory criteria 253
 late graft failure 272
 long-term monitoring 271–272
 pre-operative assessment 271
- Liver X-receptor (LXR) 718, 720
- Lobule, hepatic 232–233, 233*f*
- Locomotor system
 hyperthyroidism manifestations 389
 hypothyroidism manifestations 395
- Locus, definition 872
- Locus coeruleus, in stress response 403–404, 405
- Lofepiramine
 overdose 804
 therapeutic drug monitoring 780
- Long-acting insulin analogues 311
- Long bones 604
- Long chain hydroxyl-acyl-CoA dehydrogenase deficiency (LCHADD) 465–466, 473
- Loop of Henle 125, 125*f*
 function 128–129, 128*f*
- Lorenzo's oil 465
- Loss of heterozygosity (LOH)
 definition 872
 in tumours 868, 868*f*
- Low birth weight. *See* Birth weight, low
- Low density lipoprotein (LDL) 706–707, 708
 characteristics 708*t*
 metabolism 714
 in nephrotic syndrome 160, 160*t*
 oxidation, in atherogenesis 741–743, 742*f*
 in pregnancy 446
 receptor. *See* Low density lipoprotein receptor (LDLR)
- Low density lipoprotein (LDL) cholesterol
 effect of hormone replacement therapy 449
 effect of menopause 449
 effect of oral contraceptives 448
 measurement 732–733
- Low density lipoprotein receptor (LDLR) 714, 718–719, 719*f*; 866
 gene mutations, familial hypercholesterolaemia 727, 866
 inducible degrader of 720
- Low density lipoprotein receptor-related protein 719
- Low-dose dexamethasone suppression test 363–364
- Lower reference nutrient intake (LRNI), definition 181*t*
- Low molecular weight protein markers, renal tubular damage 162, 163*t*
- Low osmotic load hyponatraemia 49–50
- LSD (lysergic acid diethylamide) abuse 801*t*, 802
- L-type amino acid transporters (LATs) 377
- Lumbar puncture 660
 traumatic 661
- Lumbosacral radiculoplexus neuropathy 691
- Luminal digestion
 carbohydrates 221
 triglycerides 224
- Lung(s)
 disease, effects on oxygen uptake into blood 89
 function 87–88. *See also* entries beginning *respiratory*
- Lung cancer, tumour markers 836–837, 836*t*
- Luteinizing hormone (LH) 351–352, 436
 changes in men with cirrhosis 262, 262*t*
 effect of oestrogens 439
 fluctuations during menstrual cycle 435–436, 435*f*
 measurement in infertile women 440–441
 ovulation test 440
 in puberty 414
 pulsatility 436
 assessment 357
 receptor defects 425, 426*t*, 429
 secretion, assessment
 females 353
 males 352–353
 spermatogenesis role 262
 stress response 405
 testosterone secretion and 452
- Luteinizing hormone releasing hormone (LHRH). *See* Gonadotrophin-releasing hormone (GnRH)
- 17,20-Lyase 437–438
- 17,20-Lyase deficiency 425, 426*t*
- Lymphadenopathy 562
- Lymph nodes 562
- Lymphocytes 563*t*, 566
 B. *See* B lymphocytes
 count
 raised 503–504
 reference ranges 498*t*, 563*t*
 functional assays 602–603
 morphological features 504*f*
 T. *See* T lymphocytes
- Lymphocytic hypophysitis 367
- Lymphocytosis 503–504
- Lymphoid malignancies 588–597
 B cells. *See* B cell malignancies
- Lymphoid tissue 562, 562*t*, 563*f*
- Lymphoma 588, 595–596
 non-Hodgkin 506–507, 591*t*, 595
- Lysergic acid diethylamide (LSD) abuse 801*t*, 802
- Lysine
 dietary 183
 metabolism 70
 structure 170*f*
- Lysine vasopressin 42
- Lysosomal acid lipase (LAL) 717
- Lysosomal storage disorders 475, 475*b*
 substrate deprivation therapy 483
- Lysosomes, enzyme screening 475
- Lysozyme
 clearance 154*t*
 as marker of tubular function 162
- Lysylpyridinoline 611–612, 612*f*
- M**
- M2 antibodies 256
- Macroamylasaemia 217–218
- Macroprolactin 437
- Macroprolactinaemia 362, 437
- Macroprolactinoma 360–361
 assessment of remaining pituitary function 361
 management 361
- Macula densa 28–29, 126
- Maculopathy, diabetic 325
- Magnesium
 body content and distribution 117, 117*t*
 body fluid composition 28*t*
 depletion
 renal potassium loss 56. *See also* Hypomagnesaemia
 dietary intake 117
 homeostasis 117–118, 118*f*
- intestinal absorption 117, 118
 metabolism 117–121
 in chronic kidney disease 144
 poisoning, forensic biochemistry 877–878, 877*b*
 renal tubular reabsorption 117–118
 assessment 119–120, 123
 retention test 119, 123
 supplementation 121, 121*b*
 inhibition of calcium stone formation 177
 urinary excretion 117, 118
 measurement 119
- Magnesium, plasma concentration 117, 117*t*, 877–878
 analytical goals 11*t*
 analytical variation 11*t*
 biological variation 11*t*
 in chronic kidney disease 144
 critical difference 15*t*
 measurement, diagnostic approach to hypomagnesaemia 119
 in poisoning 877–878
- Magnetic resonance imaging (MRI)
 acute kidney injury 138
 evaluation of internal anatomy in disorder of sex development 418
 pituitary gland 360
- Major histocompatibility complex (MHC) 569
 locus 569*f*. *See also* Human leukocyte antigen (HLA)
- Malabsorption 219–225
 carbohydrates 222
 investigation 222
 clinical features 219–220, 220*t*
 fat 224–225
 following bariatric surgery 206
 laboratory investigations 220*t*
 proteins 223
 vitamin B₁₂ 523
- Malaria 503
 blood film 503, 503*f*
 haemolytic anaemia 528
- Maldigestion 219–225
 clinical features 220
- Male(s)
 hypogonadism. *See* Hypogonadism, male
 puberty
 delayed 420, 430
 endocrinology 414
 physical signs 415*f*, 416
 precocious 428, 429
 reproductive function 451–460
 evaluation 454–455. *See also* Spermatogenesis; Testes
 sex development 412–414, 413*f*; 451–452
 sex hormone changes in cirrhosis 262, 262*t*
 sexual dysfunction in chronic kidney disease 145
- Male infertility
 assisted reproduction techniques 457
 in chronic kidney disease 145–146
- Malignancy
 cerebrospinal fluid examination 670
 endocrine sequelae of tumours and their treatment 818–820, 819*t*
 haematological. *See* Haematological malignancies
 haematological sequelae 816, 816*t*
 hypercalcaemia of 102–103, 104, 814
 hyperuricaemia in 816
 lymphoid 588–597
 B cells. *See* B cell malignancies
 metabolic changes 813–818
 paraneoplastic syndromes. *See* Paraneoplastic syndromes; Cancer

- Malignant hyperpyrexia 654
Malignant hypertension 764
Malnutrition 200–201
 in adults 200, 201
 assessment
 hepatic secretory proteins 194–195
 immune response 195
 in children 200, 201
 hypophosphataemia 115
 nutrition support, in at-risk individuals 210*b*
 protein–energy. *See* Protein–energy malnutrition (PEM)
Malnutrition Universal Screening Tool (MUST) 193–194
Maltase 221
Maltose 221
Maltotriose 221
Management, demand on laboratories 25
Management, disease 3–4. *See also specific diseases*
Manganese 191
 deficiency 191, 199
 dietary sources 191
 laboratory-based assessment 199
 toxicity 191, 199
Mannose binding lectin (MBL) 570
Mannose therapy, congenital disorder of glycosylation 477, 477*f*
Maple syrup urine disease 467, 472
Marasmus 201
March haemoglobinuria 529
Mason-type diabetes. *See* Maturity onset diabetes of the young (MODY)
Mass spectrometry (MS), acylcarnitine analysis 657, 658*f*
Maternal complications, diabetes in pregnancy 331*b*
Maternal drug abuse 485
Maternal well-being, monitoring 445
Matrix metalloproteinases (MMPs) 743
Maturity onset diabetes of the young (MODY) 292–293
 genetic factors 287, 292–293
 MODY 1 292–293
 MODY 2 277, 292–293
 MODY 3 293
 MODY 4 293
 MODY 5 293
McCune–Albright syndrome 428–429
MDMA
 (3,4-methylenedioxymethamphetamine)
 abuse 801–802, 801*t*
MDRD (Modification of Diet in Renal Disease) 12, 134
Mean cell haemoglobin (MCH) 498
Mean cell haemoglobin concentration (MCHC) 498
Mean cell volume (MCV) 498
Mechanical ventilation 92
 shocked patients 410
Meconium aspiration 486
MedicAlert Foundation 540
Medicines and Healthcare Regulatory Authority (MHRA), blood transfusion regulation 514
Medium chain acyl-CoA dehydrogenase deficiency (MCADD) 461, 463, 464, 657
 mutation analysis 476
 neonatal screening 466
 urinary organic acid analysis 473
Medullary erythropoiesis 515
Medullary interstitial cells 127
Medullary thyroid carcinoma
 calcitonin as biomarker 400, 812, 842
 in multiple endocrine neoplasia type 2 400, 812
 treatment 813
Megalin 154
Megaloblastic anaemia 520–524
 morphological features 501, 502*f*, 520. *See also* Folate deficiency; Vitamin B₁₂ deficiency
Megestrol acetate 818
Meglitinides 315
 adverse effects 315
 hypoglycaemia 340–341
Meiosis 849
 definition 872
Melanocortins 203
 α -Melanocyte-stimulating hormone, stress response 404
Melanoma, tumour markers 837
MELAS (myopathy, encephalopathy, lactic acidosis and stroke) 656
Membrane attack complex 571*t*
Membrane-stabilizing activity, drugs 803, 804*f*
Membranoproliferative glomerulonephritis 158, 159
Membranous nephropathy 158, 159
Menaquinone 186, 186*f*
Menarche 416, 434
 premature isolated 429
Mendel, Gregor 844–845
Menin 101, 811
Meningitis, cerebrospinal fluid investigations 668, 668*t*, 669
Menkes disease 190–191
Menopause
 metabolic effects 449. *See also* Postmenopausal women
Menorrhagia, in hypothyroidism 395
Menses, onset of. *See* Menarche
Menstrual cycle 434
 effect on test results 9
 aldosterone:renin ratio 762
 hormone fluctuations 435–436, 435*f*
 hyperthyroidism manifestations 389
 sodium retention and 38
 uterine changes 436
 variation in bone turnover markers and 613. *See also* Amenorrhoea
Mental illness. *See* Psychiatric disorders
Mercury, tubular damage 161
MERRF (myoclonus, epilepsy with ragged red fibres) 656
Mesangial cells 125
Messenger RNA (mRNA) 846, 847*f*
 definition 872
Metabolic acidosis. *See* Acidosis, non-respiratory
Metabolic bone disease 604–635
 in children 632, 632*t*
 following renal transplantation 628
 genetic 632–634, 633*t*
 osteomalacia. *See* Osteomalacia
 osteoporosis. *See* Osteoporosis
 Paget disease of bone. *See* Paget disease of bone
 parenteral nutrition-associated 212–213
 in primary hyperparathyroidism 628–629
 clinical, biochemical and histological features 628–629
 treatment 629
 renal. *See* Chronic kidney disease–mineral and bone disorder (CKD–MBD)
Metabolic disease
 bone. *See* Metabolic bone disease
 inherited. *See* Inherited metabolic diseases (IMDs)
Metabolic load, reduction 481
Metabolic response, stress. *See* Stress response
Metabolic syndrome 759
 as cardiovascular disease risk factor 754, 755*f*
 definitions 754, 755*t*
 depression and 679
 prevalence 754
 type 2 diabetes association 289–290, 679
Metadrenalines, measurement, phaeochromocytoma diagnosis 763–764, 763*t*
 factors affecting results 763–764, 764*t*
Meta-iodobenzylguanidine (MIBG) scintigraphy 764
Metalloproteinases 244
Metallothionein synthesis 190
Metaphysis 604
Metformin 313–314
 contraindications 313, 314*t*
 hypoglycaemia risk 341
 mechanism of action 313
 side effects 314
 lactic acidosis 78, 313–314
 vitamin B₁₂ deficiency 523
Methadone
 abuse 801*t*
 therapeutic drug monitoring 785
 withdrawal, in neonates 485
Methaalbumin 525, 526*f*, 530
Methaemoglobin 788–791
Methaemoglobinaemia 790*t*, 805–806
 management 794*t*, 806
Methanol 685, 800
 antidote 794*t*, 800
 poisoning 685, 792*t*, 800, 801*f*
Methionine, metabolism 70
Methotrexate, therapeutic drug monitoring 783
Methylamphetamine 801
Methylcellulose 205
Methylcobalamin 521, 523*f*
3,4-Methylenedioxymethamphetamine (MDMA) abuse 801–802, 801*t*
O⁶-Methylguanine DNA methyltransferase (MGMT) 869
Methylmalonic acidemia 479, 480*f*
5-Methyl tetrahydrofolate 521, 522*f*
Meticillin-induced acute interstitial nephritis, salt wasting and 34
Metyrapone 352, 678–679
Metyrapone test 355
Microalbuminuria 152, 155
 definition 164
 in diabetes 322–323
 as marker of risk 164–166
 non-renal conditions associated with 166, 167*b*. *See also* Proteinuria
Microangiopathic haemolytic anaemia 501–503, 503*f*, 528–529
 in malignant disease 816
Microarrays 856–857
Microflora, intestinal 218–219, 561
 α_1 -Microglobulin
 as marker of tubular damage 162, 162*t*
 tubular reabsorption 154
 urinary excretion 160*t*
 β_2 -Microglobulin 594, 823*t*
 cerebrospinal fluid 667
 as marker of tubular damage 162, 162*t*
 plasma 135
 urinary excretion 160, 160*t*
Micronutrients 184
 parenteral nutrition 211. *See also* Trace elements; Vitamin(s)

- Microprolactinoma 360–361
 Microsatellites 859, 859*f*
 definition 872
 Microscopic polyangiitis 639*t*
 Microsomal triglyceride transfer protein (MTP) 711–712, 720
 defective 724
 deficiency 700
 Mid-arm circumference (MAC)
 measurements 194
 Midnight salivary cortisol 364
 Mifepristone 439
 Miglitol 317
 Miglustat 483
 Milk, breast. *See* Breast milk
 Milk-alkali syndrome 103*t*
 Milwaukee shoulder 642
 Mineral(s) 190
 balance
 in hyperthyroidism 389
 in hypothyroidism 395. *See also*
 Electrolytes; Trace elements. *specific minerals*
 Mineralocorticoid(s)
 replacement therapy, monitoring 371
 secretion 352
 Mineralocorticoid activity
 non-respiratory alkalosis 82–83
 renal tubular acidosis, type 4 173
 Mineralocorticoid excess 37*b*, 38
 hypokalaemia 55–56
 non-respiratory alkalosis 84
 syndrome of apparent excess 760*t*, 761*t*
 Minimal change disease 157–158, 159
 Minimum inhibitory concentration (MIC)
 781
 Minisatellite, definition 872
 Missense mutations 848, 848*f*
 Mitochondria
 fatty acid transport into 647–649, 649*f*
 oxidation reactions 647–649, 649*f*
 measurement 655–656. *See also* Fatty acid oxidation
 Mitochondrial disorders
 investigation. *see under* Mitochondrial respiratory chain
 neuropathy in 696, 696*t*
 Mitochondrial DNA (mtDNA) 656
 inheritance 467–468, 468*f*, 850
 mutations 656, 848–849
 inherited metabolic disease 465, 467–468
 Mitochondrial encephalomyopathy 696
 Mitochondrial isoenzyme of aspartate aminotransferase (mAST) 241
 Mitochondrial respiratory chain
 components 650*f*
 defects, investigation 655–656
 measurement of individual respiratory chain complex activity 656
 measurement of mitochondrial oxidations 655–656
 molecular biology techniques 656
 Mitogen-activated protein kinases (MAPKs)
 407
 Mitosis 849
 definition 872
 Mixed acid–base disorders 86–87
 causes 86*b*
 interpretation of acid–base data 85, 85*f*, 86
 Mixed meal test 338
 Model for End Stage Liver Disease (MELD)
 scoring system 248
 Modification of Diet in Renal Disease (MDRD) 12, 134
 Molar pregnancy 835
 Molecular chaperones, potential congenital nephrogenic diabetes insipidus therapies 43
 Molecular clinical biochemistry 2, 844–873
 applications of DNA analysis. *See* DNA analysis, applications
 glossary 871
 techniques 844
 investigation of respiratory chain defects 656. *See also* Genetic analysis, techniques.
See also entries beginning gene/genetic
 Molecular sieving theory 153–154
 Moles, invasive 835
 Molybdenum 191
 deficiency 191
 high dietary intake 191
 laboratory-based assessment 199
 Monoamine oxidase inhibitors 675
 overdose 804
 Monocarboxylate transporter 8 (MCT8)
 377
 Monoclonal gammopathy of unknown significance (MGUS) 591*t*, 596, 692
 Monoclonal proteins 568, 588–594, 589*t*
 clinical significance 588–589, 588*t*, 589*t*
 identification 589–591, 590*f*, 591*f*
 laboratory investigation 589, 589*b*
 prevalence 589
 quantitation 592–593
 transient 596
 tumour marker characteristics 822*t*
 typing 591–592, 591*t*, 592*f*
 Monocytes 563*t*, 566
 count
 raised 503–504
 reference ranges 498*t*, 563*t*
 morphological features 504*f*
 Monoglyceride(s) 704
 absorption 224, 711
 Monoglyceride lipase (MGL) 713
 Mono-iodotyrosine (MIT) 376
 Mononeuropathies 324
 Monosaccharides 221
 absorption 221, 221*f*
 Monosodium urate (MSU) crystals 639, 641*f*
 Mood disorders 675, 675*t*. *See also specific mood disorders*
 Morphine
 abuse 801*t*
 therapeutic drug monitoring 785
 Moulds, food contaminants 190
 Mouth 215
 Movement disorders 697–699
 Mucopolysaccharidoses 474, 474*t*
 Mucopolysaccharidosis type I (Hurler syndrome) 462, 474*t*
 Mucosa, innate immune system 561
 Mucosal-associated lymphoid tissue (MALT)
 562
 Müllerian ducts 413, 451
 development disorders 425
 Müllerian inhibiting substance. *See* Anti-Müllerian hormone (AMH)
 Multidrug resistance (MDR) proteins, bilirubin and bile salt metabolism 237–238
 Multifocal neuropathies 691
 Multi-organ dysfunction syndrome (MODS)
 409, 409*b*
 Multiple endocrine neoplasia (MEN)
 810–813, 810*b*
 genetic analysis 869
 type 1 216, 226–227, 810, 810*b*
 adrenal tumours 811
 diagnosis 811–812
 foregut carcinoid tumours 811
 gastroenteropancreatic neuroendocrine tumours 811
 genetics 869
 genetic screening 812
 parathyroid disease 101, 102*t*, 810–811
 pituitary tumours 811
 surveillance of patients and carriers 812, 812*t*
 tumorigenesis 811
 type 2 226–227, 810, 810*b*, 812–813
 clinical features 812, 812*f*
 diagnosis 812–813
 genetics 869
 imaging 813
 medullary thyroid cancer 400, 812
 MEN2A 810*b*, 812
 MEN2B 810*b*, 812, 812*f*
 parathyroid disease 101, 102*t*, 813
 surveillance 813
 treatment 813
 type 4 813
 Multiple myeloma 507, 591*t*, 594, 594*t*, 692
 Multiple sclerosis (MS) 671–672, 701
 Multiplex ligation-dependent probe amplification (MLPA) analysis, dystrophin gene 864–865, 865*f*
 Multiplex PCR. *See* Polymerase chain reaction (PCR)
 Muscle
 biopsy 651
 cardiac. *See* Cardiac muscle
 fibres 646, 646*f*
 structure 647*f*
 types 647*t*
 functional anatomy and physiology 646–649
 pain. *See* Myalgia
 phosphate uptake, increased 116. *See also* Skeletal muscle
 Muscle disease 646–659
 classification 650–651, 651*b*
 investigation 650–651
 biochemical 651–653
 clinical evaluation 651
 metabolic, genetically determined myopathies 654–658
 non-metabolic, genetically determined myopathies 653–654
 Muscle mass
 assessment
 functional tests 194
 laboratory-based 196
 plasma creatinine concentration and 133
 Muscle weakness 651
 in hyperthyroidism 389
 in hypothyroidism 395
 Muscular dystrophy 651*b*
 Becker. *See* Becker muscular dystrophy
 DNA analysis 864–865, 864*f*, 865*f*
 Duchenne. *See* Duchenne muscular dystrophy
 facioscapulohumeral 653
 gene mapping 845*f*
 investigation 653
 Mushrooms, poisonous 806, 806*t*
 Mutagenesis, insertional 870
 Mutation, gene. *See* Gene mutations
 Mutation analysis 848–849, 855–860
 detection of known mutations 855–857, 856*f*
 in inherited metabolic disease 476
 next generation sequencing 859–860
 scanning/screening methods 857–858, 857*f*
 tracking of mutant genes 858–859, 858*f*, 859*f*

- Myalgia 650–651
 statin-induced 653
 Myasthenia gravis 389
 Mycophenolate mofetil 600*t*, 770
 Mycophenolic acid (MPA), therapeutic drug monitoring 784–785
 Myelin-associated glycoprotein (MAG) 692
 Myelodysplasia (MDS) 506, 507*f*, 524
 Myelofibrosis 506, 506*f*
 Myeloma 507, 591*t*, 594, 594*t*, 692
 Myelopathy, hepatic 689
 Myeloproliferative neoplasms 506
 Myocardial damage, acute 744–749
 biomarkers 744–749, 745*f*
 copeptin 748
 creatine kinase-MB 744, 745*f*, 747–748
 heart-type fatty acid binding protein 745*f*, 748
 myoglobin 745*f*, 748
 post-mortem biochemistry 880
 troponins. *See* Troponin(s)
 Myocardial infarction 738, 739*f*
 definition 739, 739*b*, 744
 in diabetic patients 297–298, 330
 non-ST elevation. *See* Non-ST elevation myocardial infarction (NSTEMI)
 role of laboratory 739
 ST elevation. *See* ST elevation myocardial infarction (STEMI); *See also* Myocardial damage, acute
 Myocardial ischaemia 738, 744. *See also* Myocardial damage, acute
 Myocardium, contraction 740–741
 Myoclonus 697, 699
 Myogenic reflex 152
 Myoglobin 162, 740–741, 748
 as biomarker of acute myocardial damage 745*f*, 748
 clearance 154*t*
 Myoglobinuria 162–163, 163*t*, 653
 Myopathy. *See* Muscle disease
 Myosin 646, 647*f*, 740
 Myotonic dystrophies, investigation 654
 Myxoedema 36
 primary 396
 Myxoedema coma 397
 'Myxoedema madness' 678
- N**
 NAG (N-acetyl β -D-glucosaminidase) 161–162, 162*b*
 NAGS (N-acetylglutamate synthetase) deficiency 463*f*, 478
 Naloxone 802
 NAPQI (N-acetyl-*p*-benzoquinoneimine) 795, 795*f*
 Nateglinide 315
 Natriuresis 34
 chronic dilutional hyponatraemia 50
 Natriuretic peptides (NPs) 29–30, 30*f*
 atrial. *See* Atrial natriuretic peptide (ANP)
 cerebral salt wasting 50
 as heart failure biomarkers 749–750
 clinical utility 749–750, 750*f*
 critical values 749
 factors influencing 749
 Natural killer (NK) cells 563*t*, 566
 Nausea, arginine vasopressin secretion and 31
 Negative feedback, pituitary hormone secretion 352
 Negative predictive value 17*t*
 Nelson syndrome 678–679
 Neomycin 781
 Neonatal hepatitis syndrome 247
 Neonatal jaundice 247, 493–494
 Neonatal period, definition 484
 Neonatal screening 4, 466
 aminoacidurias 172
 congenital hypothyroidism 466, 861
 cystic fibrosis 217–218
 DNA analysis 861
 medium chain acyl-CoA dehydrogenase deficiency (MCADD) 466
 phenylketonuria 19, 861
 sickle cell disease 557–558
 Neonatal severe hyperparathyroidism 102, 102*t*
 Neonates
 with ambiguous genitalia
 evaluation 417–418, 418*f*
 general examination 418–419
 investigations 418–419. *See also* Disorder of sex development (DSD)
 bilirubin metabolism 493
 body water 487, 487*f*
 calcium and phosphorus metabolism 490–493, 490*t*
 disorders 490–493. *See also specific disorders*
 carbohydrate metabolism 488–490, 489*f*
 haemolytic disease of the newborn (HDN) 512–513
 hyperbilirubinaemia, conjugated, pathological causes 494–495, 494*b*
 hyperbilirubinaemia, unconjugated pathological causes 494, 494*b*
 physiological jaundice 247, 493–494
 hypercalcaemia 490–491
 hypernatraemia 487–488
 hyperthyroidism 387
 hypocalcaemia 108, 491, 491*b*
 hypoglycaemia 489–490, 490*b*
 hyponatraemia 487
 immune response 575–576, 576*t*
 inherited metabolic disease presentation 462–464
 defects in synthesis and breakdown 462
 enzyme deficiency disorders 463
 intoxications 462–463
 seizure disorders 463–464, 464*t*
 inherited metabolic disease screening 466. *See also* Neonatal screening
 liver disorders 493–496
 in inherited metabolic diseases 470, 471*b*
 lupus syndrome 587
 mortality rates 484
 poisoning 788
 premature. *See* Pre-term neonates
 renal function 486–488
 tests, interpretation 488
 small for gestational age 484–485
 thyroid function 379
 tumour markers 837–838
 vitamin K administration 186–187. *See also* Infants
 Neopterin 667–668
 Nephrogenic diabetes insipidus (NDI) 39
 causes 39–40, 40*b*, 42
 congenital 40–41, 43
 management 42
 water deprivation test 42
 Nephrogenic systemic fibrosis 138
 Nephrolithiasis 174
 X-linked 172. *See also* Renal calculi
 Nephrons 124–125, 125*f*
 development 486–487
 Nephropathy
 diabetic. *See* Diabetic nephropathy
 immunoglobulin A 158
 membranous 158, 159
 salt-losing 34, 34*b*
 Nephrotic syndrome 156–160
 causes 157*t*
 definition 157
 hyperlipidaemia in 730
 sodium excess 37
 Nephrotoxicity, aminoglycoside antibiotics 161, 781
 Nervous system
 effect of acidosis 76. *See also* Central nervous system (CNS)
 Nesidioblastosis 343
 Neuroblastoma 838
 Neuroendocrine tumours 808–810
 gastroenteropancreatic, in multiple endocrine neoplasia type 1 811
 gastrointestinal tract 226–227
 pancreas 226–227. *See also* Carcinoid tumours; Multiple endocrine neoplasia (MEN)
 Neurofibromatosis type 1 (NF1) 813, 868
 Neuroglycopenia 320
 acute 335, 335*b*
 chronic 335, 335*b*
 subacute 335, 335*b*
 Neurohypophysis. *See* Pituitary gland, posterior lobe
 Neurological disease 683–701
 ataxia. *See* Ataxia
 encephalopathy. *See* Encephalopathy
 inflammatory disorders of central nervous system 701
 movement disorders 697–699
 peripheral neuropathy. *See* Peripheral neuropathies
 spinal cord disorders 688–689
 causes 688*b*. *See also specific disorders*
 Neurological paraneoplastic syndromes 814
 Neuromuscular hyperexcitability, alkalosis 83
 Neuromyelitis optica (NMO) 30
 Neuronal ceroid lipofuscinosis 701
 Neuron-specific endolase (NSE)
 in cerebrospinal fluid 667, 670–671
 as tumour marker 823*t*
 lung cancer 836*t*, 837
 Neuropathy, peripheral. *See* Peripheral neuropathies
 Neuropeptide Y (NPY) 203
 Neurophysin II 31
 Neuroses 674
 Neurotensin, ectopic secretion 815*t*
 Neurotic disorders 675*t*
 Neutropenia 504
 Neutrophilia 503–504
 Neutrophils 563*t*, 564–565
 count
 low 504
 raised 503–504
 reference ranges 498*t*, 563*t*
 function tests 603
 morphological features 504*f*
 toxic changes 504, 505*f*
 pseudo-Pelger 506, 507*f*
 urine 131
 Newborns. *See* Neonates
 Niacin 188
 deficiency 694
 Nicotinamide 188
 deficiency in carcinoid syndrome 808–809
 laboratory-based assessment 198
 Nicotinic acid 188
 Niemann–Pick C1-like 1 protein 710
 Niemann–Pick disease, type C 270
 Nitisinone (NTBC) 481–482, 482*f*

- Nitrogen
balance, assessment 196
enteral feeding 210
total body, measurement 196
- Nitrogenous waste product retention, in
chronic kidney disease 143
- Nitroprusside therapy, cyanide toxicity 805
- Nitrous oxide exposure, vitamin B₁₂
deficiency association 523
- Nocturia 43
in chronic kidney disease 147
- Nocturnal polyuria 43–44
causes 43*b*
laboratory investigation 43–44, 44*f*
treatment 43–44
- Nocturnal tumescence testing 459
- Non-alcoholic fatty liver disease (NAFLD)
258
diabetes association 298
dyslipidaemia association 731
- Non-alcoholic steatohepatitis (NASH) 258
- Non-collagenous proteins, bone 606
- Non-esterified fatty acids (NEFAs). *See* Free
fatty acids
- Non-Hodgkin lymphoma 506–507, 591*t*, 595
- Non-insulinoma pancreatogenous
hypoglycaemia syndrome (NIPHS)
343
- Non-islet cell tumour hypoglycaemia
(NICTH) 282, 343–344
- Non-nucleoside reverse transcriptase
inhibitors (NNRTIs), therapeutic
drug monitoring 782
- Non-respiratory acidosis. *See* Acidosis,
non-respiratory
- Non-respiratory alkalosis. *See* Alkalosis,
non-respiratory
- Non-seminomatous germ cell tumours
(NSGCT) 832, 833*t*, 834
- Nonsense mutations 481, 848, 848*f*
- Non-small cell lung cancer (NSCLC) 836,
836*t*, 837
- Non-starch polysaccharides (NSPs) 192
- Non-ST elevation acute coronary syndrome
(NSTEACS). *See* Acute coronary
syndrome (ACS)
- Non-ST elevation myocardial infarction
(NSTEMI) 738, 744
treatment 744
- Non-steroidal anti-inflammatory drugs
(NSAIDs)
rheumatoid arthritis management 637
tubular damage 161
- Noradrenaline, stress response 405, 407–408
- Normal ranges 13–14, 13*f*
- Normoglycaemia, maintenance of 273–274,
274*b*, 275*f*
- Normoglycaemic ketoacidosis 326–327
- Northern blot, definition 872
- Nortriptyline, therapeutic drug monitoring
780
- N-telopeptide (NTX) assay 612–613, 612*f*
in children 632
- Nuchal translucency (NT) sign 443–444
- Nuclear lipids 706
- Nuclear matrix protein 22 test 827
- Nucleic acids 845. *See also* DNA
(deoxyribonucleic acid); RNA
- Nucleoside reverse transcriptase inhibitors
(NRTIs), therapeutic drug monitoring
782
- Nucleosomes 846
- Nucleotide metabolism, disorders 528
- Nucleus, antigen distribution 587, 587*f*
- Nucleus of the tractus solitarius (NTS),
appetite control and 203, 204
- Nutrition 180–199
'correct' intake 180–181
in diabetes mellitus 296, 306
requirements. *See* Nutritional
requirements
- Nutritional deficiencies
as cause of anaemia 517–520
depression and 679
peripheral neuropathy and 693–694. *See*
also specific types
- Nutritional disorders 200–213
eating disorders 207–208
malnutrition. *See* Malnutrition
management
dietary supplements 208–209
nutraceuticals 208–209
nutrition support. *See* Nutrition support
therapeutic diets 208–209, 209*t*. *See also*
Dietary management; Nutritional
management
obesity. *See* Obesity. *See also specific disorders*
- Nutritional management
acute kidney injury 140
chronic kidney disease 148. *See also* Dietary
management
- Nutritional requirements 180–192
energy. *See* Energy, requirements
micronutrients 184. *See also* Trace
elements; Vitamin(s)
protein 183
terminology 181*t*
- Nutritional status, assessment 192–199
anthropometric measurements 193–194
clinical assessment 193
dietary assessment 193
functional assessment 194–195
laboratory-based assessment of individual
nutrients 195–199
- Nutrition support 209–213
enteral feeding 210
indications 209–210, 210*b*
inherited metabolic disease 483
parenteral. *See* Parenteral nutrition
- O**
- Obesity 201–206
aetiology 202
assessment 193, 193*t*, 201–202
as cardiovascular disease risk factor
753–754
complications 202*t*
effect on test results 8
epidemiology 202
genetics 202
management 205–206
non-surgical 205
surgical. *See* Bariatric surgery
therapeutic diets 209*t*
pleiotropic syndromes 202, 202*t*
secondary causes 203
type 2 diabetes risk 286, 286*f*
- Observed results, comparison with reference
limits 15
- Obstructive acute kidney injury 137*b*, 139
- Octreotide 342–343
- Odanacatib 619
- Oedema
cerebral, in diabetic ketoacidosis 329
idiopathic 38
laboratory investigation 38
in nephrotic syndrome 159
in pregnancy 37–38
sodium excess with 36, 37, 37*b*
- Oesophagitis, reflux 214
- Oesophagus 215
- Oestradiol 438
in men 262
plasma concentration, men with cirrhosis
262, 262*t*
secretion, in puberty 414
- Oestrogen(s)
actions 439
biosynthesis 437–438, 438*f*
as cause of hypercalcaemia 103*t*
effect on biochemical variables 8*t*
effect on breast cancer 820
hormone replacement therapy 439, 447–448
hyperlipidaemia association 731
in men 262
metabolic effects 448
oral contraceptives 439, 447–448, 447*t*
plasma concentrations
fluctuations during menstrual cycle 9,
435–436, 435*f*
in men with cirrhosis 262
structure 437, 438*f*
supplementation in acute porphyrias 540
transport and metabolism 438–439
- Oestrogen-binding receptors 823*t*, 824*t*, 828
- Oestrogen deficiency
effect on calcium/bone metabolism 99
in postmenopausal women 449
management. *See* Hormone replacement
therapy (HRT)
in young women, causes 615, 615*b*
- Oestrogen replacement therapy 371. *See*
also Hormone replacement therapy
(HRT), postmenopausal women
- Oestrone 438
- Olanzapine 681, 682
- Oleander 798, 806, 806*t*
- Oligoclonal bands 664–665, 665*f*, 669–670, 671
- Oligomenorrhoea 439–440
- Oligosaccharides 221
- Oligospermia, in cirrhosis 262
- Oliguria 44, 129
in acute kidney injury 136
in acute tubular necrosis, pathogenesis
138, 139*f*
- Oncogenes 867–869
definition 872
- Oncogenic osteomalacia 112–113, 622–623
- Oncotic pressure 28
- Oocytes 433, 434*f*
assisted conception techniques 441
- Ophthalmology of Graves disease. *See* Graves
disease
- Opiates, therapeutic drug monitoring 785
- Opioids
poisoning 685, 792*t*, 802
therapeutic drug monitoring 785
- Opioid syndrome 790*t*
- Oppenheim dystonia 698
- Opsonin 571*t*
- Opsonization 565
- Optic chiasm 350, 350*f*
- Oral contraceptives (OCs) 447–450
hyperlipidaemia association 731
metabolic effects 448–449
vascular disease risk 448
- Oral glucose tolerance test (OGTT) 282,
283*b*, 300
acromegaly, monitoring response to
therapy 363
gestational diabetes screening 292*b*
- Oral hypoglycaemic agents
as cause of hypoglycaemia 346
surreptitious administration 341
screening, hypoglycaemia investigation
339. *See also* Glucose-lowering therapy,
diabetes. *specific hypoglycaemic agents*

- Orchidometry 416
 ORD (P450 oxidoreductase deficiency) 420, 423, 424*t*, 425, 426*t*
 Organic acids, urinary analysis 472–473, 473*b*
 Organic anion transporting polypeptides (OATPs) 377
 Organic disease, psychiatric manifestations 675*t*, 676–680
 Organophosphate poisoning 799
 clinical features 799
 management 794*t*, 800
 mechanisms 799
 Organ support, shocked patients 410–411
 Organ system dysfunction 409, 409*b*
 Organ transplantation 599, 599*t*
 immunological complications 599, 599*b*
 immunosuppressive therapy. *See* Immunosuppressive therapy
 inherited metabolic disease management 479–480. *See also specific organs*
 Orlistat 205
 Ornithine, structure 170*f*
 Ornithine transcarbamylase (OTC)
 deficiency 463*f*, 465, 466, 480
 Orotic acid, urinary analysis 473
 Orthopaedic implants, blood metal concentrations and 799
 Orthostatic proteinuria 155
 Osmolal gap
 calculation 62
 in poisoning 875*t*
 transurethral prostatectomy syndrome 46
 Osmolality
 definition 27
 plasma 27, 30
 measurement in poisoning 792*t*
 serum, calculation 62
 urine. *See* Urine osmolality
 Osmoles, idiogenic 45
 Osmoreceptor response 31
 Osmoregulation
 in pregnancy 41
 water intake 31–32
 water output 30–31, 31*f*
 Osmotic demyelination syndrome 51
 Osmotic fragility tests 530
 Osmotic pressure 27
 Osteitis fibrosa 628
 Osteoarthritis (OA) 637
 basic calcium phosphate crystals and 642
 Osteoblasts 96, 605, 607, 608
 defective function, osteomalacia and 623–624
 Osteocalcin 606
 age-related changes 610*f*, 632
 as bone turnover marker 610–611, 610*f*
 in Paget disease of bone 630
 treatment response 631–632
 Osteoclasts 605, 608
 effect of calcitonin 98
 maturation 96
 in Paget disease 630
 Osteocytes 605, 607–608
 Osteocytogenesis 607–608
 Osteodystrophy, renal. *See* Chronic kidney disease–mineral and bone disorder (CKD–MBD)
 Osteogenesis imperfecta 632–634, 633*t*
 Osteolysis, familial expansile 633*t*, 634
 Osteolytic metastases, hypercalcaemia and 102–103
 Osteomalacia 620–624
 acidosis and 623
 calciopenic 620–622
 causes 620*b*
 laboratory investigation 621–622, 621*t*
 responses to therapy 622
 defective osteoblast function 623–624
 oncogenic 112–113, 622–623
 phosphopenic 116, 622–623, 622*b*
 laboratory investigation 621*t*, 623
 treatment 623
 tumour-induced 816
 Osteonectin 607
 Osteopenia
 definition 613–614
 diabetes association 298
 of prematurity 491–492, 492*t*
 Osteopetrosis 633*t*, 634
 Osteophytes 637
 Osteopontin 607, 607*t*
 Osteoporosis 613–620, 614*f*
 causes 614–615, 615*b*
 definition 613–614, 616
 hormone replacement therapy and 450, 618
 investigation and diagnosis 615–617
 biochemical investigation 616–617, 616*t*
 bone densitometry 615–616
 risk factors 614, 614*b*
 treatment 617–620
 biochemical responses 619–620, 619*f*
 calcium and vitamin D 617
 lifestyle modifications 617
 pharmacological management 617–619
 Osteoporotic fractures 614
 clinical risk factors 615
 estimation of absolute risk 616
 Osteoprotegerin 96, 608, 613
 Osteosarcoma 630
 Osteosclerosis 633*t*, 634
 Otorrhoea, cerebrospinal fluid 666
 Ovarian cancer 838–840
 detection of residual disease 839
 diagnosis 839
 long-term surveillance 840
 monitoring 839–840
 prognosis 839
 risk of malignancy index (RMI) 839, 839*t*
 screening 838–839
 Ovarian failure
 primary, as cause of amenorrhoea 440
 in Turner syndrome 431
 Ovarian hyperstimulation syndrome (OHSS) 441
 Ovarian steroid hormones 437–439
 actions 439
 biosynthetic enzymes and pathways 437–438, 438*f*
 secretion through menstrual cycle 435–436, 435*f*, 438
 structure 437, 437*f*
 transport and metabolism 438–439. *See also specific hormones*
 Ovaries 433–435, 434*f*, 435*f*
 chemotherapy effects 819–820
 development 413–414, 413*f*
 polycystic. *See* Polycystic ovary syndrome (PCOS)
 Overnight dexamethasone suppression test 364
 Ovulation 434, 435
 onset of 414
 tests 440
 Oxalate
 foods rich in 177*b*, 178–179
 intestinal absorption, increased 176
 urinary excretion, increased 175, 176–177
 Oxalic acid crystals 639, 643
 Oxcarbazepine, therapeutic drug monitoring 778
 Oxidation, fatty acids. *See* Fatty acid oxidation
 Oxidative phosphorylation 68
 Oxidative stress, diabetic neuropathy pathophysiology 692
 Oximeters 91
 3-Oxoacyl-CoA thiolase 649*f*
 measurement of activity 657–658
 2-Oxoglutarate 71, 72*f*
 17-Oxosteroid reductase 438
 Oxygen
 fraction of inspired gas comprising (FiO₂) 87
 partial pressure in arterial blood. *See* PaO₂ (partial pressure of oxygen in arterial blood)
 Oxygen delivery
 tissues 89, 91, 91*f*
 measurement 90–91. *See also* Tissue oxygenation
 Oxygen–haemoglobin dissociation curve 88–89, 88*f*, 90, 90*f*, 550
 normal *vs.* sickle haemoglobin 551*f*
 Oxygen therapy, respiratory distress in neonates 486
 Oxygen transport
 role of haemoglobin 88–89, 88*f*, 517, 550
 to tissues 89–90. *See also* Tissue oxygenation
 Oxygen uptake
 into blood 87–88
 effects of pulmonary disease 89
 into tissues 90, 90*f*. *See also* Tissue oxygenation
 Oxyhaemoglobin, in cerebrospinal fluid 661, 666
 Oxytomodulin (OXM) 205
 Oxyterol 715
 Oxytocin 352
 labour role 447
 pituitary function assessment 354
P
 p53 868
 P450. *See* Cytochrome P450
 PaCO₂ 87
 effects of pulmonary disease 89
 laboratory assessment 73. *See also* PCO₂ (partial pressure of carbon dioxide)
 Paediatrics 484–496. *See also* Children; Infants; Neonates
 Paget disease of bone 629–632
 aetiology 629
 clinical features 630
 epidemiology 629
 investigations 630–631
 biochemical tests 630–631
 radiology 630
 juvenile 633*t*, 634
 natural history 629–630
 pathology 630
 treatment responses 631–632, 631*f*
 Pain
 abdominal, in acute porphyria 538, 541
 in articular disorders 638–639
 bone, pagetic 630
 chronic, acute porphyria complication 539
 muscle. *See* Myalgia
 Pancreas 217–218
 autoimmune disease 583, 601*t*
 neuroendocrine tumours 226–227
 transplantation 318
 Pancreatic cancer, tumour marker 840
 Pancreatic enzymes, serum measurement 217–218

- Pancreatic function tests 217–218
 choice for pancreatitis 229
 direct/invasive 217
 non-invasive 217–218
- Pancreatic juice, fluid composition 34*t*
- Pancreatic lipase-related proteins 717
- Pancreatic polypeptide (PP) 203, 204
- Pancreatic triglyceride lipase (PTL) 716, 717
- Pancreatic β -Cells
 deficiency/dysfunction in type 2 diabetes 286–288
 destruction in type 1 diabetes 284–285
 measurement of function 303
- Pancreatitis 217
 acute 228–229
 as cause of diabetes mellitus 293
 causes 228, 229*t*
 choice of test for 229
 chronic 217
 as cause of diabetes mellitus 293
 investigation. *See* Pancreatic function tests
- Pancreozymin 217
- Pancytopenia 504
- Panhypopituitarism 360
- Panic disorder 676–677
- Pantothenic acid 189
 deficiency 694
 laboratory-based assessment 198
- PaO_2 (partial pressure of oxygen in arterial blood) 87, 88
 effects of pulmonary disease 87, 89
- Papaverine 459
- Papillary thyroid cancer 399
- Paracetamol
 metabolism 795, 795*f*
 overdose 792*t*, 795–806
 acute liver failure 253
 clinical features 795
 management 794*t*, 795–796, 796*f*
 toxic dose 795
- Paraganglioma 763
- Paralysis, periodic. *See* Periodic paralysis
- Paraneoplastic syndromes 814–816
 humoral 814
 neurological 696–697, 697*t*, 814
- Paraproteinaemia 163–164, 163*t*, 588–589, 589*t*
 in myeloma 594
 transient 596. *See also* Monoclonal proteins
- Paraproteinaemic neuropathies 692–693
- Paraquat poisoning 792*t*
- Parasellar tumours 367
- Parasuicidal gestures 788
- Parathyroid disease
 carcinoma 101, 102*t*
 in multiple endocrine neoplasia type 1 101, 102*t*, 810–811
 in multiple endocrine neoplasia type 2 101, 102*t*, 813
- Parathyroidectomy 627–628, 629
 hypocalcaemia and 108, 627, 629
- Parathyroid hormone (PTH) 95–96, 95*f*
 actions on kidneys 96
 in chronic kidney disease 144
 in chronic kidney disease–mineral and bone disorder 625, 626
 diurnal variation 95
 effect of magnesium concentrations 118
 effect on TmP/GFR 111
 excess. *See* Hyperparathyroidism
 in hypercalcaemia 175
 measurement
 chronic kidney disease–mineral and bone disorder investigation 626
 circulating PTH 96
 during/after parathyroidectomy 629
 hypercalcaemia investigation 103–104, 105*f*
 hypocalcaemia investigation 109
 receptors 95–96
 resistance in pseudohypoparathyroidism 106–107, 107*t*
- Parathyroid hormone–calcitriol–FGF23 axis, in chronic kidney disease–mineral and bone disorder 624–625
- Parathyroid hormone-related peptide (PTHrP) 99
 ectopic secretion 814, 815*t*
 effect on TmP/GFR 111
 hypercalcaemia of malignancy and 102
 investigation of hypercalcaemia 104
- Parathyroid hormone (PTH) therapy
 biochemical responses 619
 osteoporosis 618
- Paraventricular nucleus (PVN), appetite control and 203, 204
- Parenteral nutrition 210, 211–213
 complications 211–213
 composition of fluids 211
 liver function test abnormalities and 264
 monitoring 211, 212*t*
 prolonged, as cause of cholestatic liver disease 495
 in short bowel syndrome 213
- Parkinson disease 697
- Parkinsonism 697–699
- Paroxetine
 overdose 804
 therapeutic drug monitoring 780
- Paroxysmal nocturnal haemoglobinuria (PNH) 524, 529
 diagnosis 531
- Partial androgen insensitivity syndrome (PAIS) 419–420, 426*t*, 427–428, 456
- Patent ductus arteriosus (PDA) 486
- Patient access, results 23
- Patient education, diabetes mellitus 307
- Paul–Bunnell antibody 500
- PCO_2 (partial pressure of carbon dioxide) 67, 73
 laboratory assessment 73
 lungs, alveolar *vs.* venous 67
 in non-respiratory alkalosis 82
 relationship with hydrogen ion concentration 85, 85*f*
 in respiratory acidosis 81
 restoration to normal 81–82
 in respiratory alkalosis 84
- PCR. *See* Polymerase chain reaction (PCR)
- Pellagra 188
- Pemphigoid 585
- Pemphigus 585
- Pendred syndrome 396
- Pendrin 375–376, 375*f*
- Penetrance, definition 851, 872
- Penicillamine challenge test 267
- Penicillamine therapy
 adverse reactions 267
 cystinuria 171, 171*f*
 Wilson disease 267
- Penicillins, renal potassium loss 56
- Penile arterial blood pressure, measurement 459
- Penile prosthesis 459
- Pentamidine, hypoglycaemia association 346
- Pentose phosphate pathway 517, 518*f*
 disorders 526–528
- Pentose phosphate shunt 278
- Pepsin 223
- Peptide YY (PYY) 203, 204
- Peptidyl disulfide isomerase (PDI) 605
- Perchlorate 375
- Perchlorate discharge test 388
- Percutaneous endoscopic gastrostomy (PEG) 210
- Performance indicators, laboratory medicine 24–25, 24*t*
- Periodic paralysis
 hyperkalaemic 59, 654
 hypokalaemic 52–53, 654
- Periodontal disease 208
- Peripheral blood, as stem cell source 600, 600*t*
- Peripheral blood stem cell transplant (PBSCT) 600
- Peripheral cyanosis 88
- Peripheral nervous system
 hyperthyroidism manifestations 389
 hypothyroidism manifestations 394–395
- Peripheral neuropathies 689–697
 acute inflammatory neuropathies 692
 bariatric surgery-associated 694
 causes 690*b*
 chronic inflammatory demyelinating polyneuropathies 692–693
 in chronic kidney disease 693
 diabetic. *See* Diabetic neuropathy
 in endocrine disturbances 693
 immune-mediated 692
 investigations 690, 690*b*
 in liver disease 693
 metabolic 690*b*, 694–696
 in mitochondrial disorders 696
 nutritional 693–694
 paraneoplastic 696–697, 697*t*
 paraproteinaemic 692–693
 porphyric 538, 695
 signs and symptoms 690
 small fibre painful axonal neuropathy 691
 symmetrical 690
 causes 690*t*
- Peripheral oedema 36
- Periportal hepatitis 254
- Peritoneal dialysis
 acute kidney injury management 141
 chronic kidney disease management 150
 continuous ambulatory. *See* Continuous ambulatory peritoneal dialysis (CAPD)
 hyperkalaemia management 61
 inherited metabolic disease management 483
 thyroid hormone concentrations and 146
- Peritonitis, ascites and 264–265, 265*t*
- Perls reaction 525, 530
- Pernicious anaemia 217, 523, 583
 antibody tests 523
- Peroxisomal disorders, investigation 474
- Peroxisome proliferator-activated receptor family 315, 720
- Peroxisome proliferator-activated receptor γ analogues. *See* Thiazolidinediones
- Persistent hyperinsulinaemic hypoglycaemia of infancy (PHHI) 490
- Personnel, laboratory 22
- Petrosal sinus sampling,
 adrenocorticotrophic hormone 365–366
- Peyer's patches 562
- pH 66
 blood 66
 urine 70–71, 131
- Phaeochromocytoma 677, 763–764
 biochemical investigation 763–764, 763*t*
 as cause of hypercalcaemia 103*t*
 inherited 763, 813
 localization 764
 management 764

- in multiple endocrine neoplasia type 2 813
prevalence 763
secondary diabetes 294
symptoms 763
- Phagocytes
deficiency
 associated infections 576, 577*t*
 primary immunodeficiency 579
dysfunction in diabetes 297. *See also*
 Monocytes; Neutrophils
- Phagocytosis 565
- Phagosome 565
- Pharmacodynamic(s) 767, 768*f*, 769
definition 767
monitoring 776–777
- Pharmacogenetics 682, 776, 861–862
- Pharmacogenomics 776, 861–862
- Pharmacokinetics 767–769, 768*f*
definition 767
- Phenformin-induced lactic acidosis 78
- Phenobarbital
poisoning 792*t*
therapeutic drug monitoring 779
- Phenobarbitone, effect on biochemical variables 8*t*
- Phenotype, definition 872
- Phenylalanine hydroxylase deficiency 479*t*,
851. *See also* Phenylketonuria (PKU)
- Phenylketonuria (PKU) 461, 462
heterogeneity 851
incidence 472
inheritance 466–467
management 481
maternal 466
neonatal screening 19, 861
- Phenytoin
effect on biochemical variables 8*t*
poisoning 792*t*
therapeutic drug monitoring 770–771,
774, 779, 779*f*
- PHEX** 112–113
- Phobic anxiety disorders 676–677
- Phosphate
acute deficiency syndrome 116, 116*t*
body fluid composition 28*t*
buffering 67
dietary 111, 114, 115
 in chronic kidney disease 148–149
fluxes 110*f*
homeostasis 109, 110*f*
intestinal absorption 110*f*, 111
intracellular 110
renal tubular handling 110*f*, 111–113, 169
 disorders 172. *See also* TmP/GFR
 (tubular maximum for phosphate/
 glomerular filtration rate)
renal tubular loss, osteomalacia/rickets
 resulting from 622, 622*b*
replacement
 diabetic ketoacidosis management 328
 intravenous 117
supplementation, in phosphopenic
 osteomalacia 623
triple stones 177
urinary excretion, increased 116. *See also*
 Phosphorus
- Phosphate, plasma concentration 113, 113*t*
analytical goals 11*t*
analytical variation 11*t*
biological variation 11*t*
in chronic kidney disease 144
in chronic kidney disease–mineral and
 bone disorder 626
critical difference 15*t*
diurnal variation 100*f*, 113
in osteopenia of prematurity 492, 492*t*
- Phosphate metabolism 109–117
in chronic kidney disease 144
control, hyperphosphataemia management
 627
in neonates. *See* Neonates
renal, disorders 112–113
- Phosphatonins 111, 112–113
- Phosphocreatine 646–647
- Phosphodiesterase-5 (PDE-5) inhibitors 459
- Phosphofructokinase 278, 654–655
- 3-Phosphoglycerate dehydrogenase
deficiency 477–478
- Phosphoglycerate kinase 654–655
- Phosphoglycerate mutase 654–655
- Phospholipase A2 (PLA2) 717–718
- Phospholipids 704
digestion and absorption 224
function 703*t*
nuclear 706
structure 705*f*
- Phospholipid transfer protein (PTP) 718
- Phosphopenic osteomalacia. *See*
 Osteomalacia
- Phosphoric acid, daily production and
elimination 68*t*
- Phosphorus
blood 109, 109*b*
distribution, body 109–113, 109*b*
intracellular 109–110
metabolism. *See* Phosphate metabolism;
 Phosphate
- Phosphorylase a 654–655
- Phosphorylase kinase 654–655
- Phosphorylated tau 665
- Photocoagulation, diabetic retinopathy
treatment 325–326
- Photosensitivity
porphyrin-induced 543
protoporphyrin-induced 546
treatment 547
- Physical activity. *See* Exercise
- Phytanic acid, accumulation 694
- Phytanoyl-CoA hydroxylase deficiency 694
- Phytomenadione 186, 186*f*
- Phytosterolaemia 711
- Phytosterols 704
- Pick disease 670–671
- Piecemeal necrosis 253, 254
- Pigment epithelium-derived factor (PEDF)
633
- Pioglitazone 315, 316
contraindications 316
- Piper methysticum* (kava) 208–209
- Pituitary adenomas 360
ACTH-secreting. *See*
 Adrenocorticotrophic hormone
 (ACTH)-dependent Cushing
 syndrome
 gonadotrophin-secreting 367
 growth hormone-secreting. *See*
 Acromegaly
in multiple endocrine neoplasia type 1 811
non-functioning 367
prolactin-secreting. *See* Prolactinoma
thyroid stimulating hormone-secreting
 366–367
- Pituitary apoplexy 367
- Pituitary disease
clinical assessment 360
deficiency states 367–369. *See also*
 Hypopituitarism
diagnostic techniques 360. *See also*
 Pituitary gland, assessment of
 function
hormone replacement therapy, monitoring
 370–371
- hypersecretion states 360–367
monitoring of pituitary function 359–360.
 See also specific pituitary diseases
- Pituitary fibrosis, idiopathic 367
- Pituitary gland
adenomas. *See* Pituitary adenomas
anatomy 350, 350*f*
anterior lobe 350, 350*f*
 physiology 350–351, 351*f*
assessment of function 352–360, 371–372
 basal hormonal investigations 352–354,
 371
 clomifene test 356–357
 in disease states 359–360. *See also specific*
 pituitary diseases
dynamic tests of ACTH–adrenal
 function 354–355
gonadotrophin secretion 352–353,
 356–357
growth hormone reserve 355–356
luteinizing hormone pulsatility 357
posterior pituitary 357
protocols 357–358, 371–372
releasing hormone tests 356. *See*
 also Cortisol, pituitary function
 assessment. *specific tests*
deficiency states 367–369
imaging
 in Cushing disease 366
 techniques 360
insufficiency, hypoglycaemia due to 345
irradiation
 acromegaly management 363
 pituitary function assessment following
 359–360
posterior lobe 350, 350*f*
 function tests 357
 physiology 351*f*, 352
surgery. *See* Pituitary surgery
thyroid hormone uptake 377
tumours 367
 adenomas. *See* Pituitary adenomas
 craniopharyngiomas 367
- Pituitary–gonadal axis 351–352
assessment 352–353
- Pituitary surgery
acromegaly management 363
Cushing syndrome management 366
reassessment following 366
pituitary function assessment following
 359
- Pituitary–thyroid axis 351
assessment 352
- Placental alkaline phosphatase (PLAP) 823*t*,
832
- Placental site trophoblastic tumours 835
- Plant sterols 757
- Plant toxins 806, 806*t*
- Plaque, dental 208
- Plasma bile acids (PBA), measurement 243
- Plasma cells 588
- Plasmacytoma, solitary 595
- Plasma protein A, pregnancy-associated
443–444
- Plasma proteins 663*t*
buffering capacity 67
liver function tests. *See* Liver function tests
pregnancy-related changes 446. *See also*
 Protein(s). *specific plasma proteins*
- Plasma renin activity (PRA). *See* Renin
- Plasma viscosity, measurement 499
- Plasma volume, pregnancy-related changes
445
- Plasmin 507
- Plasminogen activator inhibitor-1 (PAI-1),
cardiovascular disease risk and 753

- Plasminogen kringle 4 710
 Plasminogen kringle 5 710
Plasmodium falciparum malaria 503, 503*f*
 haemolytic anaemia 528
 Platelet(s)
 cardiovascular disease association 753
 count 498*t*, 499
 haemostasis role 507, 515
 transfusion 513
 Platelet-activating factor acetylhydrolase (PAF-AH) 717–718
 Platelet transfusion, in severe sepsis 410
 Pleiotropic obesity syndromes 202, 202*t*
 Podocytes 153, 153*f*
 POEMS syndrome 693
 Poikilocytosis 501
 Point mutations 846–847
 definition 872
 Point-of-care testing 12, 25–26
 advantages 25
 areas of concern 26
 therapeutic drug monitoring 775
 Poisoning 787–807
 accidental 788
 adults 788
 aetiology 788
 antidotes 794, 794*t*
 body temperature and 793
 central nervous system complications 791–793
 childhood 788
 definition 787
 diagnosis 789–791
 with endogenous agents. *See* Forensic biochemistry
 factors involved 788*f*
 indications for measurement of drugs and poisons 792*t*
 infants 788
 intrauterine 788
 management 791–795
 methaemoglobinaemia. *See* Methaemoglobinaemia
 neonates 788
 renal complications 793
 specific poisons 795–806
 alcohol (ethanol) 685, 792*t*, 799–800, 801*f*
 antidepressants 803–804
 benzodiazepines. *See* Benzodiazepine overdose
 carbon monoxide. *See* Carbon monoxide poisoning
 chloroquine 798
 chromium 799
 cobalt 799
 cyanide. *See* Cyanide poisoning
 digoxin 798
 drug and substance abuse 800–803
 ethylene glycol. *See* Ethylene glycol
 iron. *See* Iron poisoning
 lead 792*t*, 798–799
 lithium 792*t*, 804–805
 methanol. *See* Methanol
 organophosphates. *See* Organophosphate poisoning
 paracetamol. *See* Paracetamol
 plant and fungal toxins 806, 806*t*
 salicylate. *See* Aspirin
 theophylline. *See* Theophylline
 syndromes caused by (toxidromes) 789, 790*t*
 types of lesion 788–789, 789*t*
 Poisons centres 791, 807
 Polarography 656
 Polycystic ovary syndrome (PCOS)
 diagnostic criteria 440
 hirsutism 441
 hyperprolactinaemia 361
 management 440
 oligomenorrhoea/amenorrhoea 439
 ovarian hyperstimulation syndrome risk 441
 secondary diabetes 294
 ultrasonographic images 435*f*
 Polydipsia
 primary
 management 42
 polyuria secondary to 41
 psychogenic 41, 681
 secondary, primary polyuria with 39–41
 Polyethylene glycol (PEG)-uricase 642
 Polygenic, definition 872
 Polygenic hypercholesterolaemia 728
 Polyglandular syndrome, autoimmune 106
 Polymerase chain reaction (PCR) 852–855, 854*f*
 definition 872
 in investigation of CNS infections 668, 669
 multiplex 856–857, 856*f*
 definition 872
 in mutation analysis 855, 856*f*
 cystic fibrosis 863–864, 864*f*
 dystrophin gene 864–865, 864*f*
 in prenatal diagnosis 860
 real-time 855
 reverse transcriptase 855
 Polymeric feeds 210
 Polymorphic markers, linkage analysis 858, 859*f*
 Polymorphisms 848
 balanced, definition 871
 definition 872
 drug metabolising enzymes 861
 single-stranded conformational 857, 857*f*
 Polymorphs. *See* Neutrophils
 Polymyositis 639*t*
 Polyol pathway, diabetic neuropathy pathophysiology 691
 Polyostotic fibrous dysplasia 633*t*, 634
 Polysaccharides, non-starch 192
 Polystyrene sulphonate resins 61
 Polyunsaturated fatty acids 183*f*, 224. *See also* Fatty acid(s)
 Polyuria 39–43, 129
 definition 39
 laboratory investigation 41–42
 management 42–43
 nocturnal. *See* Nocturnal polyuria
 pregnancy and 41
 primary, with secondary polydipsia 39–41
 secondary to primary polydipsia 41
 Pontine myelinolysis 51
 Population screening 4, 861
 Porphobilinogen (PBG)
 in acute porphyria 230
 synthesis 533, 535*f*
 urinary, measurement 539
 Porphyria(s) 533–549, 534*t*
 abbreviations 535*b*
 acute 230, 534, 534*t*, 538–542
 anaesthesia and 542
 autosomal dominant 538–542
 chronic complications 539
 clinical presentation 538–539
 diagnosis 536*f*, 537*t*, 539
 homozygous 542
 management 539–540
 managing asymptomatic relatives 541, 541*t*
 pathophysiology 538
 pregnancy and 542
 prevention 540
 rare forms 542
 safe prescribing 541
 severely affected patients 540–541
 acute intermittent. *See* Acute intermittent porphyria (AIP)
 biochemical findings 537*t*
 clinical manifestations 534
 congenital erythropoietic. *See* Congenital erythropoietic porphyria (CEP)
 cutaneous 534*t*, 542–547
 bullous. *See* Bullous porphyrias
 erythropoietic protoporphyria. *See* Erythropoietic protoporphyria (EPP)
 X-linked dominant protoporphyria. *See* X-linked dominant protoporphyria (XLDPP)
 hepatic 268
 hepatoerythropoietic 542–543, 545
 inheritance 467
 laboratory investigations 536*f*
 molecular genetics 534–538, 537*t*
 overview 534
 variegate. *See* Variegate porphyria (VP)
 Porphyria cutanea tarda (PCT) 258, 534*t*, 544–545
 biochemical findings 537*t*
 molecular genetics 537*t*, 544
 treatment 544
 Porphyric neuropathy 538, 695
 Porphyrin-induced photosensitivity 543
 Porphyrinogens 533–534, 535*f*
 Porphyrins 533–534
 adult reference ranges 537*t*
 laboratory investigations 533–534, 536*f*, 537*t*, 539
 metabolism, disorders
 secondary 547–548, 548*t*. *See also* Porphyria
 Portal hypertension 259
 Portal vein 232, 232*f*
 Positional cloning, definition 872
 Positive predictive value 17*t*
 Postanalytical factors, affecting test results 12–13
 Post-heparin lipolytic activity (PHLA) 716, 733
 Post-hypercapnic alkalosis 81, 83
 Postmenopausal women
 bisphosphonate use 618
 calcium intake 617
 cardiovascular disease risk 751–752
 hormone replacement therapy. *See* Hormone replacement therapy (HRT)
 ovarian cancer screening, CA125 measurement 838–839
 Post-mortem biochemistry 874, 875*b*, 878–880
 interpretation 879*t*
 little used/unvalidated tests 879–880, 879*t*
 specific diagnostic problems 880–882
 anaphylaxis/anaphylactic reactions 880
 diabetes 878*b*, 880, 881*f*
 drowning 880
 hypothermia/hypothermia 880–882
 inflammation 882
 sudden death 882, 882*t*
 vitreous humour 879–880, 879*t*
 Post-obstructive natriuresis 34
 Postpartum thyroiditis 399
 Postprandial hypoglycaemia 347
 Postprandial syndrome 347–348
 Postrenal acute kidney injury 137*b*, 139
 Post-traumatic stress disorder 679, 680
 Posture, effect on test results 8*t*, 9
 aldosterone:renin ratio 762
 urinary protein excretion 155

- Potassium
 body fluid composition 28*t*
 concentrations
 plasma. *See* Potassium, plasma
 concentration
 stool water 34*t*, 53–54
 urine 33
 vitreous humour, post-mortem 879
 content of oral preparations 57, 57*t*
 depletion. *See* Potassium depletion
 dietary intake 57
 excess. *See* Hyperkalaemia
 extracellular fluid and 28*t*, 32–33
 extrarenal fluid composition 34*t*
 homeostasis, effect of acidosis 76
 infusion 57
 intracellular fluid and 28*t*, 32–33
 metabolism
 in chronic kidney disease 143
 disorders 52–61. *See also* Hyperkalaemia;
 Hypokalaemia
 renal control of output 33
 renal tubular handling 169*t*
 retention 59, 59*b*
 in synovial fluid, post-mortem 879–880
 Potassium, plasma concentration 32
 analytical goals 11*t*
 analytical variation 11*t*
 biological variation 11*t*
 critical difference 15*t*
 effect on aldosterone:renin ratio 762
 in hyperkalaemia 57–58
 in hypokalaemia 52
 in poisoning 875*t*
 Potassium bicarbonate 57, 57*t*
 Potassium chloride 57, 57*t*
 Potassium citrate
 hypercalciuria management 176
 hyperoxaluria management 176–177
 hypokalaemia management 57, 57*t*
 inhibition of calcium stone formation 177
 Potassium depletion 33
 in diabetic ketoacidosis 55, 328
 extrarenal causes 53–54, 53*b*
 non-respiratory alkalosis 82–83
 renal causes 54, 54*b*. *See also* Hypokalaemia
 Potassium salts 57, 57*t*
 Potassium-sparing diuretics, potassium
 retention 59
 Prader orchidometer 416
 Prader-Willi syndrome 202*t*, 852
 Pralidoxime 799
 Pravastatin 653
 Prealbumin. *See* Transthyretin
 Preanalytical factors, affecting test results
 7–10
 Precision, analytical 11, 11*t*
 Precocious puberty 428
 causes 428, 428*b*
 central (gonadotrophin-dependent)
 428–429, 428*b*
 endocrine investigations 428*t*
 variants of early puberty 429
 Precursor supply, inherited metabolic disease
 management 476–477
 Predictive value of tests 3, 16–20
 definitions 16–17, 16*t*, 17*t*
 disease prevalence and 18, 18*t*
 example 17–18, 17*t*, 18*f*
 practical application 19
 Pre-digested feeds 210
 Prednisolone therapy
 autoimmune hepatitis 256
 effect on cortisol assays 358
 non-islet cell tumour hypoglycaemia
 344
 Pre-eclampsia 38, 445, 446, 764
 Pregnancy 442–447
 acute fatty liver of 263, 445, 465–466
 antinuclear antibodies in 587
 biochemical changes 445–446
 biochemical diagnosis 442
 biochemical monitoring 443
 carbon monoxide poisoning in 788
 cholestasis of 262–263, 445
 diabetes mellitus
 complications 331*b*
 management 331–332
 type 1 291–292, 331–332
 drug abuse during 485
 ectopic. *See* Ectopic pregnancy
 hypertension in 38, 156, 764
 inherited metabolic disease presentation
 465–466
 intrauterine infections 484–485, 485*t*
 liver function during 262–263
 molar 835
 polyuria and 41
 porphyria patients 542
 proteinuria in 156, 446
 screening for fetal malformation. *see under*
 Fetus
 sodium retention 37–38
 thyroid function 379–380, 385, 445
 urinary protein excretion 156. *See also*
entries beginning gestational; Labour
 Pregnenolone 437, 438*f*
 Preimplantation genetic diagnosis 444, 444*t*
 haemoglobinopathies 558–559
 Prematurity 484
 apnoea of 486
 Premises, laboratory 22
 Prenatal diagnosis 444, 860–861
 inherited metabolic diseases 444, 476
 Preproinsulin 278–279
 Proliferative retinopathy, diabetic 325
 Prerenal acute kidney injury. *See* Acute
 kidney injury (AKI)
 Prescribing, safe, in acute porphyria 541
 Pre-term labour 447
 Pre-term neonates 484
 apnoea of prematurity 486
 body water 487, 487*f*
 osteopenia of prematurity 491–492, 492*t*.
See also Birth weight, low
 Prevalence, disease
 predictive value of tests and 17*t*, 18. *See also*
specific diseases
 Priapism, in dialysed males 146
 Primary aldosteronism (PA) 760–762. *See also*
 Hyperaldosteronism
 Primary biliary cirrhosis (PBC) 256, 639*t*
 Primary myxoedema 396
 Primary sclerosing cholangitis (PSC)
 256–257
 Primidone, therapeutic drug monitoring 779
 Principal cells 126
 Probability 16, 19–20
 Probe
 definition 872
 hydridization 852–855
 Probiotic 209
 Procainamide, therapeutic drug monitoring
 778
 Procalcitonin 99
 as marker of acute phase response 598
 Procollagen 605, 606*f*
 Procollagen I amino-terminal extension
 peptide (PINP) 611
 Procollagen I carboxy-terminal extension
 peptide (PICP) 611
 Procollagen I extension peptides 611
 Procollagen type III (PIIINP), plasma
 measurements 244
 Product, alternate, inherited metabolic
 disease management 478
 Product breakdown, inhibition in inherited
 metabolic disease management 478
 Product replacement, inherited metabolic
 disease management 477–478
 Progastrin releasing peptide (ProGRP) 836*t*,
 837
 Progesterone
 actions 439
 biosynthesis 437, 438*f*
 plasma concentrations
 in ectopic pregnancy 442
 fluctuations during menstrual cycle 9,
 435–436, 435*f*
 ovulation assessment 440
 pituitary function assessment 353
 in spontaneous abortion 443
 structure 438*f*
 transport and metabolism 438–439
 Progesterone-binding receptors 823*t*, 824*t*,
 828
 Progesterone-only contraceptives 447*t*, 448,
 449
 Progestogens
 actions 439
 hormone replacement therapy 448
 metabolic effects 448
 oral contraceptives 439, 447*t*, 448
 structure 437
 Proglucagon 316
 Prognosis
 use of biochemical tests 3. *See also specific*
diseases
 Progression, disease, monitoring 3–4
 Progressive diaphyseal dysplasia 633*t*, 634
 Prohormone convertase 404
 Proinsulin 278–279, 281
 abnormalities 279, 280*t*
 measurement, hypoglycaemia investigation
 337*f*, 338–339
 Prolactin 437
 abnormalities in males with chronic kidney
 disease 145, 146*b*
 actions 437
 deficiency 351, 354
 ectopic secretion 815*t*
 effect of stress 9, 361
 plasma concentration
 females 440
 pituitary function assessment 352
 in prolactinoma 352
 secretion 351
 abnormalities in psychiatric disease 681
 dynamic tests 361
 modulators 437
 synthesis 437
 tumour marker characteristics 822*t*
 Prolactinoma 360–362
 management 361–362
 presentation 361–362
 Proliferative retinopathy, diabetic 325
 Prolonged fast test 336–338
 Prolyl peptidyl cis-trans isomerase B 605
 Promoter, definition 872
 Proopiomelanocortin (POMC) 203, 204,
 404, 814
 obesity and 202
 Propionic acidemia 479, 480*f*
 Propoxyphene abuse 801*t*
 Propranolol, hypoglycaemia association 346
 Proprotein convertase subtilisin kexin 9
 (PCSK9) 720–721
 gain of function gene mutations 727, 866

- Propylthiouracil 380, 391
 Prostacyclin 705, 706f
 Prostaglandin(s) 705, 706f
 labour role 447
 Prostaglandin D-synthase, cerebrospinal fluid 663f, 663t, 666
 Prostaglandin E₂ 406, 706f
 Prostaglandin endoperoxide synthase 705
 Prostaglandin G₂ 705, 706f
 Prostaglandin H₂ 705, 706f
 Prostate cancer 840–841
 androgen effects 820
 management 841
 screening and diagnosis 840–841
 Prostate cancer gene 3 protein (PCA3) 823t
 Prostate-specific antigen (PSA) 822t
 analytical and reporting requirements 826t, 841
 in prostate cancer management 25, 841
 in prostate cancer screening/diagnosis 824t, 840–841
 Protease activated receptors (PARs) 408
 Proteases 223
 Protein(s)
 absorption 222–224
 clinical aspects 223
 investigation 224
 bone matrix 605–607
 brain-specific 663f, 665–666
 cerebrospinal fluid. *See* Cerebrospinal fluid (CSF)
 conservation by kidneys 152–155
 deficiency 183
 dietary 183, 222
 dietary restriction in chronic kidney disease 148–149, 323
 digestion 223
 drug binding 769
 enteral feeding 210
 laboratory-based assessment 196
 loss, as cause of immune deficiency 579–580, 579t
 markers of tubular damage 162t
 high molecular weight 161–162
 low molecular weight 162, 163t
 metabolism. *See* Protein metabolism
 parenteral nutrition 211
 reabsorption, tubular 154
 secretion, tubular 154–155
 synthesis 235
 stress response 405
 urinary excretion 131, 155, 155t
 determinants 155–156
 normal 131, 155–156, 155t
 selectivity 158–159. *See also* Proteinuria
 urine, measurement 165–166, 166f. *See also* Plasma proteins. *specific proteins*
 Protein C 408–409
 activated 411
 Protein concentration
 analytical goals 11t
 analytical variation 11t
 biological variation 11t
 critical difference 15t
 effect of posture 9
 sodium deficiency 35
 Protein disulphide isomerase (PDI) 720
 Protein–energy malnutrition (PEM) 200
 in children 201, 223
 hypoglycaemia 347. *See also* Malnutrition
 Protein kinase C (PKC), activation, in diabetic tissue damage 297
 Protein metabolism
 in chronic kidney disease 144
 effects of alcohol 258
 hepatic 235
 stress response 405
 14-3-3 Proteins 665, 670–671
 Proteinuria 131, 152–167
 Bence Jones. *See* Bence Jones proteinuria
 clinical correlates 157t
 clinical investigation 165–167
 stepwise 166–167, 166f
 exercise-induced 155–156
 postural 155
 in pregnancy 156, 446
 of prerenal origin 162–164
 in renal disease 156–164
 glomerular. *See* Glomerular proteinuria
 staging and prognosis of chronic kidney disease 156
 tubular. *See* Tubular proteinuria
 urine protein selectivity 158–159
 Proteoglycans
 bone 606–607
 heparan sulphate 713
 Proteomics 871
 Prothrombin 507, 508f
 Prothrombin time (PT) 508, 509t
 liver function tests 242, 248t
 Proton pump inhibitors (PPIs), hypomagnesaemia and 118
 Proto-oncogenes 867
 Protoporphyrin
 erythropoietic. *See* Erythropoietic protoporphyria (EPP)
 X-linked dominant. *See* X-linked dominant protoporphyria (XLDPP)
 Protoporphyrin liver disease 547
 Protoporphyrin, in extravascular haemolysis 525
 Protoporphyrin-induced photosensitivity 546
 treatment 547
 Protoporphyrinogen oxidase (PPOX) 533–534, 535f
 Provocation tests, hypoglycaemia 336–338
 Proximal convoluted tubule 125, 125f
 amino acid reabsorption 168
 function 128
 tests 136
 glucose absorption 168
 phosphate reabsorption 169
 protein reabsorption 154
 Pseudo-Cushing syndrome 293, 800
 Pseudodementia 677
 Pseudogout 642
 Pseudohermaphroditism 416, 417t.
 See also Disorder of sex development (DSD)
 Pseudohyperkalaemia 58, 60
 Pseudohyperphosphataemia 113b, 114
 Pseudohypoaldosteronism 56, 59t, 60, 60t
 laboratory investigations 60–61
 type I 35, 56, 60–61, 60t
 type II 60–61, 60t, 760t, 761t
 Pseudohypoglycaemia 336
 clinical 320
 Pseudohyponatraemia 46
 Pseudohypoparathyroidism 106–107, 107t
 classification 107t, 122
 diagnosis 122
 in neonates 491
 Pseudo-Pelger neutrophils 506, 507f
 Pseudopseudohypoparathyroidism (PPHP) 107, 107t
 Psychiatric disorders 673–682
 aetiology 675
 classification 674–675, 675t
 compulsive water drinking 41, 681
 endocrine and metabolic manifestations 680–681
 future developments 682
 investigations 674
 biochemical 675–676
 organic disease manifestations 675t, 676–680
 prevalence 674
 Psychiatry 673–674
 investigations 674
 subspecialities 675
 Psychogenic polydipsia 41, 681
 Psychological factors, cardiovascular disease risk 754
 Psychosis 674
 Korsakoff 187, 686
 Psychotropic drugs
 metabolic complications 675–676, 681–682. *See also* Antipsychotic drugs
 Pteroylpolyglutamate hydrolase 521
 Pteroylpolyglutamic acid 521, 522f. *See also* Folate
 Puberty
 disorders of 412, 428
 endocrine investigations 428t. *See also* Delayed puberty; Precocious puberty
 inherited metabolic disease presentation 464–465
 normal 414–416
 endocrinology 414
 physical signs 414–416, 415f
 tempo of 414–415
 Pubic hair, growth of 415–416, 415f
 Pulmonary aspiration, enteral feeding-associated risk 210
 Pulmonary function 87–88
 Pulmonary hypertension, sickle cell patients 556
 Pulse oximeters 91
 Pure red cell aplasia 816
 Purine metabolism 639–640, 640f
 Pyelonephritis, urinary sediment 131
 Pyrazinamide, therapeutic drug monitoring 782
 Pyridinoline 611–612, 612f
 Pyridoxal 188, 188f
 Pyridoxal phosphate 477
 Pyridoxamine 188, 188f
 Pyridoxamine 5-phosphate oxidase deficiency 477
 Pyridoxine 188, 188f
 challenge, in homocystinuria 479
 hyperoxaluria and 176
 pharmacological doses 188. *See also* Vitamin B₆
 Pyrimidine 5'-nucleotidase 517, 518f
 deficiency 528
 tests 531
 Pyrosequencing 860
 Pyruvate 70, 277f, 278
 blood measurements in inherited metabolic diseases 473–474
 Pyruvate carboxylase deficiency 78
 Pyruvate dehydrogenase 277f, 278
 deficiency 78, 670
 Pyruvate kinase (PK) deficiency 528, 531
- Q**
 Quality, definition 21
 Quality assurance 24, 26
 external. *See* External quality assurance (EQA)
 therapeutic drug monitoring 776
 Quality control, internal. *See* Internal quality control (IQC)
 Quality indicators 24–25, 24t
 Quality management systems 22

- Quality standards 21–23
 Quetiapine 681
 Quinagolide 361, 440
 Quinidine-like activity, drugs 803, 804f
 Quinine, hypoglycaemia association 346
- R**
- Rabson–Mendenhall syndrome 281
 Race, cardiovascular disease risk and 752
 Radioiodine therapy
 as cause of hypothyroidism 396
 Graves disease 391–392
 thyroid cancer 400
 toxic adenoma 392
 toxic multinodular goitre 392
 Radiology
 chronic kidney disease–mineral and bone disorder investigation 627
 hypercalcaemia investigation 104
 Paget disease investigation 630
 thyroid gland 387
 Radionuclide techniques, assessment of renal function 130
 Radiotherapy, effects
 reproductive 819–820
 on somatic growth 818–819
 Raloxifene 618
 RANK (receptor activator of nuclear factor- κ B) 608, 629
 RANKL (receptor activator of nuclear factor- κ B ligand) 608, 609f, 613
 Rapoport–Luebering shunt 517, 518f
 Rasburicase 642, 816
 RAST (radioallergosorbent test) 581
 RB (retinoblastoma gene) 867–868
 rBAT 171
 Reactive hypoglycaemia 347
 Reactive oxygen species (ROS), in diabetic tissue damage 297
 Real-time polymerase chain reaction 855
 Reaven syndrome. *See* Metabolic syndrome
 Receiver operating characteristic (ROC) curves 19, 19f
 Recessive, definition 872
 Recombination 849, 849f
 definition 872
 Recombination fraction 849
 Recommended daily amount (RDA), definition 181t
 Recommended daily intake (RDI), definition 181t
 Rectum, villous adenoma of 53–54
 Red blood cell(s) 515–517
 carbon dioxide transport and buffering 67, 67f
 in cerebrospinal fluid 661
 counting 498
 enzyme defects 526–528, 527b
 tests 531
 formation. *See* Erythropoiesis
 function 517, 518f
 glycolytic intermediates, measurement 531
 indices 498
 in iron deficiency anaemia 520
 lysis. *See* Haemolysis
 membrane defects
 acquired 527b, 529
 inherited 525–526, 527b
 metabolites
 measurement 531. *See also specific metabolites*
 morphology. *See* Red blood cell morphology
 removal by reticuloendothelial system 517
 structure 516–517. *See also* Red blood cell morphology
 survival, measurement 530
 transfusion 513
 in severe sepsis 410
 urinary casts 131, 166
 Red blood cell morphology
 abnormalities 501, 502f, 516–517
 in anaemia 501–503
 haemolytic 529
 normal 501, 501f, 516–517
 Redistribution hyperkalaemia. *See* Hyperkalaemia
 Redistribution hypokalaemia. *See* Hypokalaemia
 Reducing substances, urinary, in inherited metabolic diseases 470–471
 5 α -Reductase deficiency 426t, 427, 456
 Refeeding syndrome 115, 207, 211–212
 identification of people at increased risk of developing 211–212, 212b
 Reference change value 14, 15
 Reference intervals 14
 disadvantages 14–15
 Reference limits 14
 comparison of observed results with 15
 Reference nutrient intake (RNI), definition 181t
 Reference values 14–15
 Reflux oesophagitis 214
 Refsum disease 694–695
 Regular insulin 310
 Regulation, laboratories 22, 22b
 Rejection, graft 599, 599b
 Releasing hormone tests 356
 Remnant hyperlipoproteinaemia (RH) 725–726, 726f
 Renal blood flow 127, 152
 pregnancy-related changes 445
 Renal bone disease. *See* Chronic kidney disease–mineral and bone disorder (CKD–MBD)
 Renal calculi 174–179
 infection-related stones 177
 investigation of stone formers 178, 178b
 pathogenesis 175–178
 rarities 177–178
 treatment 178–179
 types 175t
 uric acid stones 175t, 177
 Renal disease 129–130
 acidosis 78–80
 autoimmune 585, 585t, 586f, 601t
 diabetic. *See* Diabetic nephropathy
 hyperphosphataemia 114
 hypertension in 759
 manifestations 129
 proteinuria. *See* Proteinuria
 slowing progression of 148–149
 therapeutic diets 209t. *See also specific diseases*
 Renal failure
 acute. *See* Acute kidney injury
 in acute porphyria 539
 chronic. *See* Chronic kidney disease
 encephalopathy and 686
 established, in diabetes mellitus 323
 hypoglycaemia associated with 344–345
 peripheral neuropathy and 693
 polyuria 39
 Renal function 127–129
 assessment 130–136
 measurement of glomerular filtration rate. *See* Glomerular filtration rate (GFR)
 in neonates, test interpretation 488
 in poisoning 793
 urinalysis 130–131
 biochemical tests 130–136
 development 486–487
 glomerular function 127
 impaired
 lipid abnormalities 730–731
 renin levels 147, 762. *See also* Renal disease
 involving hormones 144, 144t
 in neonates. *See* Neonates
 in pregnancy 445
 progressive loss of 142. *See also* Chronic kidney disease (CKD)
 tubular function 128–129. *See also* Kidney(s)
 Renal hypokalaemia, without specific acid–base disorder 54b, 56
 Renal hypokalaemic acidosis 54–55, 54b, 57
 Renal hypokalaemic alkalosis 54b, 55–56, 57
 Renal leak hypercalciuria 175–176
 Renal osteodystrophy. *See* Chronic kidney disease–mineral and bone disorder (CKD–MBD)
 Renal replacement therapy
 acute kidney injury 141
 chronic kidney disease 149–150. *See also specific methods*
 Renal sodium loss
 primary 34, 34b
 secondary 35, 35b
 Renal transplantation
 bone disease following 628
 chronic kidney disease 150
 cystinuria 171
 hyperlipidaemia following 731
 Renal tubular acidosis (RTA) 54, 78–80, 172–173
 diagnosis 179
 type 1 (distal) 79–80, 173, 492
 type 2 (proximal) 79–80, 173, 492
 type 3 173
 type 4 (hyperkalaemic) 79–80, 173
 in diabetes 326
 Renal tubular disorders 168–174
 generalized tubular defects 174
 isolated abnormalities of tubular function 169–174
 osteomalacia/rickets resulting from 622, 622b. *See also* Renal tubules, damage
 Renal tubules
 damage
 assessment methods 161–162
 drug-induced 161
 heavy metal-induced 161
 high molecular weight protein markers 161–162
 low molecular weight protein markers 162, 162t. *See also* Renal tubular disorders
 development 486–487
 distal. *See* Distal convoluted tubule (DCT)
 function 128–129
 assessment 161–162
 intrinsic control
 potassium 33
 sodium 28
 physiology 168–169, 169f, 169t
 protein reabsorption 154
 protein secretion 154–155
 proximal. *See* Proximal convoluted tubule. *See also entries beginning tubular;* Kidney(s)
 Renin 28–29, 29f
 direct active renin concentration 762
 effect of posture 9
 menstrual cycle and 38
 plasma activity 762
 reference ranges 766
 in renal impairment 147, 762

- Renin (*Continued*)
 in pre-eclampsia 38
 secretion 126
 ectopic 815*t*
 in stress response 406
 in syndromes of hypoaldosteronism 59–60, 59*t*. *See also* Aldosterone:renin ratio
- Renin–angiotensin–aldosterone system
 28–29, 29*f*
 in idiopathic oedema 28
 in pregnancy 37
- Renovascular hypertension 759–760
- Repaglinide 315
- Reproductive function
 effects of malignant disease therapy 819–820
 female. *See* Female(s)
 hormones regulating 436–437. *See also specific hormones*
 male. *See* Male(s)
- Reproductive system
 hyperthyroidism manifestations 389
 hypothyroidism manifestations 395, 440
- Request forms 6, 7
- Resonium A 54
- Respiratory acidosis. *See* Acidosis, respiratory
- Respiratory alkalosis. *See* Alkalosis, respiratory
- Respiratory chain. *See* Mitochondrial respiratory chain
- Respiratory compensation
 non-respiratory acidosis 73, 74–75
 non-respiratory alkalosis 82
- Respiratory control ratio (RCR) 656
- Respiratory disorders
 effects on oxygen uptake into blood 89
 neonates 485–486
- Respiratory distress syndrome (RDS)
 485–486, 485*b*
 management 486
- Respiratory failure 89
 encephalopathy and 686–687
 management 91–92
- Respiratory quotient (RQ), calculation 195, 195*f*
- Respiratory support 92
 in poisoning 791
 respiratory distress in neonates 486
 shocked patients 410
- Respiratory system
 hyperthyroidism manifestations 389
 hypothyroidism manifestations 395
- Response-to-injury hypothesis 741
- Restriction endonuclease(s) 852–853, 853*f*
 allele recognition 855, 856*f*
 definition 872
- Restriction fragment length polymorphism (RFLP) 852, 858, 859*f*
 definition 872
- Reticulin 233–234
- Reticulocyte(s) 515–516
 count 498*t*, 499
- Reticuloendothelial system, red blood cell removal 517
- Retinal 184, 184*f*
- Retinoblastoma 867–868
- Retinoblastoma gene (*RB*) 867–868
- Retinoic acid 185
- Retinoic acid receptors (RAR) 185
- Retinoid X receptors (RXR) 185, 377, 720
- Retinol 184, 184*f*
 teratogenicity 185
- Retinol-binding protein 184
 as marker of tubular damage 162, 162*t*
 tubular reabsorption 154
- Retinopathy, diabetic. *See* Diabetic retinopathy
- RET* proto-oncogene mutations 812
- Retrovirus
 definition 872
 gene therapy vectors 870
- Reverse transcriptase 853–854
 polymerase chain reaction (PCR) 855
- Reye-like illness 496
- Reye syndrome 496
- Reynolds risk score 751
- RGD (arginyl-glycyl-aspartate)-containing glycoproteins 607, 607*t*
- Rhabdomyolysis 162
 in acute porphyria 538
 aminotransferase activities 240
 causes 162–163, 163*b*
 creatine kinase activity 652
 hyperphosphataemia and 114
 hypocalcaemia and 107–108
- Rheology, cardiovascular disease risk and 753
- Rhesus (Rh) blood group 510
 laboratory tests 511, 511*t*
- Rheumatoid arthritis (RA) 585–586, 637
 anaemia in 643–644
 management 637
- Rheumatoid factor (RF) 585, 644
- Rhinorrhoea, cerebrospinal fluid 666, 666*f*
- Rhodopsin 184–185
- Riboflavin (vitamin B₂) 187
 deficiency 187
 laboratory-based assessment 197–198
- Rickets 492, 492*b*, 620
 hypophosphataemic. *See* Hypophosphataemic rickets
 vitamin D-dependent. *See* Vitamin D-dependent rickets (VDDR)
- Riedel thyroiditis 399
- Rifampicin, therapeutic drug monitoring 782
- Rimonabant 205
- Ringer's solution 36*t*
- Risedronate 618, 631*f*
- Risk of malignancy index (RMI), ovarian cancer 839, 839*t*
- RNA (ribonucleic acid) 845
- RNA polymerase 846
- Rose Bengal dye test 247
- Rosen Waaler test 644
- Rosiglitazone 315
- Rotor syndrome 246, 548
- Roux-en-Y gastric bypass (RYGB) 206, 206*f*, 207*t*
- Rubeosis iridis 325
- Ruboxistaurin mesylate 297
- S**
- S100 family of proteins 665
 S100B as melanoma marker 837
- Safe prescribing, in acute porphyria 541
- St John's wort, drug interactions 208–209, 771, 772*t*
- Salbutamol, hyperkalaemia management 61
- Salicylates
 hypoglycaemia association 346. *See also* Aspirin
- Saline
 diabetic ketoacidosis management 328
 non-respiratory alkalosis management 83
 sodium deficiency management 36*t*
- Saline infusion, hypertonic. *See* Hypertonic saline infusion
- Saliva 215
 amylase 221
 therapeutic drug monitoring 773
- Salt intake 30. *See also* Sodium, dietary intake
- Salt-losing nephropathy 34, 34*b*
- Salt poisoning 878, 878*b*
- Salts, potassium 57, 57*t*
- Salt wasting
 cerebral 50
 meticillin-induced acute interstitial nephritis and 34
- Sampling, therapeutic drug monitoring 772–774
- Sandhoff disease 700
- Sarcoidosis 367
- Sarcomere 647*f*, 740
- Saturated fat, dietary 224
- Saturated fatty acids 704, 704*f*
- Scavenger receptor(s) 742
 class B type 1 (SRB1) 719
- Schilling test 524
- Schizophrenia 675, 679–680
- Schmidt syndrome 584*t*
- Schofield equation 196, 196*f*
- Schumm test 530
- Scintigraphy
 chronic kidney disease–mineral and bone disorder investigation 627
 meta-iodobenzylguanidine (MIBG) 764
 thyroid. *See* Thyroid scintiscanning
- Scleroderma 587, 587*t*, 639*t*
- Sclerostin 607–609, 609*f*, 613, 619
- Screening 4–5
 acute porphyrias 541*t*
 diabetes mellitus 301, 302*f*
 diabetic retinopathy 325, 325*b*
 DNA analysis 861. *See also* Genetic screening
 fetal malformation. *see under* Fetus individuals 4–5, 861
 mutation detection 857–858, 857*f*
 neonatal. *See* Neonatal screening population 4, 861
 thyroid disease 401
- Scurvy 189
- Secretin–pancreozymin test 217
- Secretory piece 568
- Sedative–hypnotic poisoning 790*t*
- Sediment, urine 131
- Seizures
 in acute porphyria 538
 neonatal, inherited metabolic disease presentation 463–464, 464*t*. *See also* Convulsions
- Selective oestrogen receptor modulators (SERMs) 618
- Selective screening 4
- Selective serotonin reuptake inhibitors (SSRIs)
 overdose 804
 therapeutic drug monitoring 780
- Selenium 191
 deficiency 191
 dietary sources 191
 laboratory-based assessment 199
 toxicity 191
- Semen analysis 454, 454*t*
- Seminiferous tubules 451, 452*f*
- Seminoma 453, 832
 diagnosis 832
 long-term surveillance 829
 prognosis 833*t*
 teratoma *vs.* 832. *See also* Germ cell tumours
- Sensitivity 16–17, 17*t*
- Separation techniques, lipoproteins 733–734
- Sepsis
 as cause of delirium 676
 definition 409*b*
 hypoglycaemia association 347

- management 410
 immunomodulation 411
 organ support 410
- Septic encephalopathy 687–688
- Septo-optic dysplasia, congenital 39–40, 367
- Serotonin 204
 appetite control 204
 foods rich in 810*b*
 metabolism 809, 809*f*
- Serotonin agonist syndrome 790*t*
- Sertoli cell(s) 451, 452*f*
 in chronic kidney disease 145–146
 tumours 453, 841–842
- Sertraline
 overdose 804
 therapeutic drug monitoring 780
- Serum enzymes, pancreatic function testing 217–218
- Serum free light chains (SFLC) 596
- Severe combined immunodeficiencies (SCID) 578
 X-linked, gene therapy 870
- Severity assessment, disease 3
- Sex. *See* Gender
- Sex assignment 416
- Sex development
 disorders. *See* Disorder of sex development (DSD)
 normal 412–414, 413*f*, 451
- Sex hormone(s)
 changes in cirrhosis 261–263
 changes in hyperthyroidism 389
 females. *See* Ovarian steroid hormones
 loss of, effect on bone density 615, 615*b*
 measurement, investigation of disorder of sex development 420, 421*f*
 physiology and biochemistry 262
 replacement therapy, monitoring 371
 stress response 405. *See also specific sex hormones*
- Sex hormone-binding globulin (SHBG) 262
- androgen sensitivity assessment 428
 in cirrhosis 262, 262*t*
 in hypothyroidism 395
 oestrogen transport 438–439
 testosterone binding 439
 in thyroid stimulating hormone-secreting adenomas 366–367
- Sexual dysfunction
 in chronic kidney disease 145–146
 in chronic liver disease 261–262
- Sexual orientation 416
- SGLT1 (sodium-dependent glucose cotransporter 1) 170, 276, 276*t*
- SGLT2 (sodium-dependent glucose cotransporter 2) 170, 276, 276*t*
- Sheehan syndrome 367
- Shock 409
 anaphylactic 574–575
- Shocked patients 409–411
 definitions 409, 409*b*
 management 410–411
 immediate care 410
 immunomodulation 411
 organ support 410–411
- Short bowel syndrome 213
- Short stature, causes 368. *See also* Growth hormone (GH) deficiency
- Short tetracosactide test 354, 355, 358, 372
- Shoulder, capsulitis 643
- Shunting 89
- Sialoprotein, bone 607, 607*t*
- Sibutramine 205
- Sickle cell syndrome 49
- Sickle euthyroid syndrome 380, 678, 680
- Sickle cell disease 525, 550, 555–556
 clinical manifestations 555–556
 epidemiology 550, 551*f*, 555
 laboratory diagnosis 557–559, 557*t*, 558*f*
 neonatal screening 557–558
 pathophysiology 555
 prognosis 556
- Sickle haemoglobin (HbS) 555, 557*t*
- Sildenafil 459
- Silent thyroiditis 398
- Single nucleotide polymorphisms (SNPs)
 in gene tracking 858–859. *See also* Point mutations
- Single-stranded conformational polymorphism (SSCP) 857, 857*f*
- Sirolimus 150, 600*t*
 therapeutic drug monitoring 784
- Sitosterol 757
- β -Sitosterolaemia 711
- Sjögren syndrome 587, 587*t*, 638, 639*t*
- Skeletal muscle
 cardiac muscle *vs.* 740
 functional anatomy and physiology 646–649
 glucose metabolism 274–275
 effects of insulin 279–280, 280*t*
 glycogen 274
 insulin actions 280*t*. *See also* Muscle
- Skeleton
 hyperthyroidism manifestations 389
 hypothyroidism manifestations 395
- Skin
 autoimmune disease 585, 601*t*
 development 488
 hyperthyroidism manifestations 389
 Graves disease 390
 hypothyroidism manifestations 395
 innate immune system 561
 porphyrias 534*t*, 542–547
 bullous. *See* Bullous porphyrias
 erythropoietic protoporphyria. *See* Erythropoietic protoporphyria (EPP)
 X-linked dominant protoporphyria. *See* X-linked dominant protoporphyria (XLDPP)
- Skinfold thickness 194
- Skin prick testing 581, 581*t*
- Small bowel
 normal microflora 218. *See also* Duodenum; Ileum; Jejunum
- Small bowel bacterial overgrowth 218–219
 causes 219
 definition 219
 diagnosis 219
 symptoms 219
- Small cell lung cancer (SCLC) 836, 836*t*
- Small fibre painful axonal neuropathy 691
- Small for gestational age (SGA) neonates 484–485
- Smith–Lemli–Opitz syndrome 462
- Smoking
 as cardiovascular disease risk factor 752
 effect on bone mineral density 615
- Smoking cessation, diabetes mellitus management 307
- Smooth muscle antibody (SMA) 255–256
- Sodium
 appetite 30
 body fluid composition 28*t*
 concentrations
 plasma. *See* Sodium, plasma concentration
 vitreous humour, post-mortem 879
 deficiency. *See* Sodium deficiency
 deficit calculation in hypovolaemic hyponatraemia 62
- dietary intake 28
 cardiovascular disease risk and 756
 in diabetes 306
 effect on aldosterone:renin excretion 762
 effect on calcium excretion 95
 hypokalaemia and 38
- excess. *See* Sodium excess
- expected depression in hyperglycaemia-induced hyponatraemia, estimation 62
- extracellular fluid 28–30, 28*t*, 128
- extrarenal fluid composition 34*t*
- fractional excretion of 878, 878*b*
- metabolism, disorders 33–39
 sodium deficiency. *See* Sodium deficiency
 sodium excess. *See* Sodium excess
- plasma concentration. *See* Sodium, plasma concentration
- renal control of output 28–30, 128
- renal tubular handling 168, 169, 169*t*
- retention
 in ascites formation 259–260
 in chronic kidney disease 147
 in hepatorenal syndrome 261
 in patients with diabetes and hypertension 289
- urinary excretion 28
 chronic kidney disease 34, 147, 147*f*
 hyponatraemia 50
 nocturnal polyuria 43–44, 44*f*
- vitreous humour, post-mortem biochemistry 879*t*
- water intoxication requirement, estimation 62
- Sodium, plasma concentration 28
 analytical goals 11*t*
 analytical variation 11*t*
 biological variation 11*t*
 critical difference 15*t*
 in hyponatraemia 46
 in poisoning 875*t*
 in polyuria 41
- Sodium bicarbonate, non-respiratory acidosis management 76
- Sodium chloride poisoning 878, 878*b*
- Sodium deficiency 33–36
 causes 33–35
 clinical presentation 33
 laboratory investigation 35–36
 management 36, 36*t*. *See also* Hyponatraemia
- Sodium-dependent glucose cotransporter 1 (SGLT1) 170, 276, 276*t*
- Sodium-dependent glucose cotransporter 2 (SGLT2) 170, 276, 276*t*
- Sodium excess 36–39
 causes 37–39, 37*b*
 clinical presentation 36
 laboratory investigation 39
 management 39
 with oedema 36, 37, 37*b*
 without oedema 37*b*, 38–39. *See also* Hypernatraemia
- Sodium–glucose cotransporter 1 (SGLT1) 170, 276, 276*t*, 333
- Sodium–glucose cotransporter 2 (SGLT2) 170, 276, 276*t*, 333
- Sodium–iodide symporter 375–376
- Sodium lactate 36*t*
- Sodium loading, acute 38
- Sodium polystyrene sulphonate 54
- Sodium valproate. *See* Valproate
- Soil-eating 54
- Solanine 189
- Solitary plasmacytoma 595

- Solvent abuse 792*t*, 803
Somatic cell, definition 872
Somatic growth
 effects of malignant disease treatments 818–819. *See also* Growth
Somatoform disorders 675*t*
Somatostatin 351
 analogues, acromegaly management 363
 ectopic secretion 815*t*
 inhibition of thyroid-stimulating hormone release 379, 379*f*
Somatostatinoma 227
Somatostatin syndrome 222
Somogyi effect 320–321
Sorbitol 296
 diabetic neuropathy pathophysiology 691–692
Sortilins 712–713, 721
Sorting protein-related receptor with type A repeats (SORLA) 721
Southern blotting 852–853
 definition 872
SOX9 412–413
Soy protein 756
Space of Disse 233–234, 233*f*, 713
Specificity 11, 16–17, 17*t*
Specimen carry-over 826*b*
Spectrin 516–517
Sperm
 count 454, 454*t*
 motility 454
Spermatogenesis 262, 414, 451
 defective
 in chronic kidney disease 145–146
 in cirrhosis 262
 regulation 453, 453*f*
Spherocytes, osmotic fragility tests 530
Spherocytosis, hereditary. *See* Hereditary spherocytosis
Sphingolipidoses 270
Sphingolipids 705
 function 703*t*
 structure 707*f*
Sphingomyelin 705, 707*f*
 cell membranes 703–704
 fetal lung production 444–445
Sphingosine 705, 707*f*
Spinal cord disorders 688–689
 causes 688*b*
Spinal osteosclerosis 628
Spinal paracentesis. *See* Lumbar puncture
Splenectomy, as cause of immune deficiency 580
Splice site mutations 848, 848*f*
Splicing 846
Spontaneous abortion 443
Spurious hyperkalaemia 58
Spurious hypernatraemia 44
Spurious hypokalaemia 52
Spurious hyponatraemia 46
SQSTM1 mutations 629
Squamous cell carcinoma (SCC) antigen 823*t*, 829
 as cervical cancer marker
 monitoring 829
 prognosis 829
 screening and diagnosis 829
 as lung cancer marker 836*t*, 837
SRV (sex-determining region Y) gene 412–413
Staff
 laboratory 22
 therapeutic drug monitoring service 775
Standard deviation 10, 15
Stanols 704
Starch 181–182, 182*f*, 220–221
 absorption 221–222, 221*f*
Starling forces 28
Starvation, metabolic changes 817*t*
Statin therapy 734, 735*t*
 adverse effects 735*t*, 736
 as cause of elevated creatine kinase activity 653, 736
 cerebrotendinous xanthomatosis 696
 in chronic kidney disease 149
 drug interactions 653, 736
 management of diabetes-related cardiovascular risk 308
 mode of action 735*t*
 nephrotic hyperlipidaemia 160
 rate of cholesterol absorption as predictor of benefit 711
Steatohepatitis, non-alcoholic 258
Steatorrhoea 224–225
Steatosis 234–235
 alcoholic 257–258
 diabetes association 298
 non-alcoholic fatty liver disease 258
ST elevation myocardial infarction (STEMI) 738, 739*f*, 744
 treatment 744
Stellate cells 233–234, 244
Stem cells 564, 564*f*
 in gene therapy 870–871
 sources 600*t*
Stem cell transplantation 600
 immunological complications 599*b*. *See also* Bone marrow transplantation
'Steroid card' 371
Steroid hormones
 measurement, investigation of disorder of sex development 420, 421*f*
 metabolism 438*f*
 ovarian. *See* Ovarian steroid hormones
 pregnancy-related changes 445
 synthesis. *See* Steroidogenesis; Sex hormone(s). *specific hormones*
Steroidogenesis 352, 353*f*
 ovarian 437–438, 438*f*
Steroid response elements (SREs) 721
Steroid therapy 600*t*
 adrenal suppression, assessment 358–359. *See also* Prednisolone therapy
Sterol regulatory element binding proteins (SREBPs) 721
Sterols 703–704
 plant 757
Stick tests, blood ketones 299–300
Stigmasterol 757
St John's wort, drug interactions 208–209, 771, 772*t*
Stomach 215–217
 cancer. *See* Gastric cancer
 emptying 204
 gastrin 215, 216
 Helicobacter pylori infection. *See Helicobacter pylori* infection
 intrinsic factor. *See* Intrinsic factor
 microflora 218
 protein digestion 223
 stretching 204. *See also entries beginning gastric*
Stones, renal. *See* Renal calculi
Stool water, composition 34*t*, 53–54
Strachan syndrome 694
Stress
 amenorrhoea and 369
 definition 403
 effect on diabetes mellitus 322
 effect on test results 8*t*, 9
 inflammation and 406–409
 Stress hyperglycaemia 273–274
 Stressors, categories 403
 Stress-related disorders 675*t*
 Stress response 403–405
 acute phase proteins 408
 adrenal medulla 405
 catecholamines 407–408
 central 403
 coagulation factors 408–409
 cytokines 406
 hypothalamo-pituitary-adrenal axis 404–405
 inflammation and 406–409
 initiation 403, 404*f*
 kidneys 406
 shock 409. *See also* Shocked patients
 systemic 403
 Stroke 737–738
 sickle cell patients 556
 Strong ion difference (SID) 85–86
 Strong ion gap (SIG) 86
 Strontium ranelate 618–619, 620
 Struma ovarii 393
 Subacute sclerosing panencephalitis (SSPE) 670–671
 Subarachnoid haemorrhage 661–662, 670
 Subclinical thyroid disease 385
 hyperthyroidism. *See* Hyperthyroidism
 hypothyroidism. *See* Hypothyroidism
 Substance abuse 800–803
 psychiatric disorders 675–676, 675*t*
 Substrate depletion, inherited metabolic disease management 483
 Substrate deprivation therapy, inherited metabolic diseases 483
 Succinate dehydrogenase 655, 656
 Succinylcholine, redistribution hyperkalaemia and 58
 Sucrase-isomaltase 221
 Sucrose 220–221
 cardiovascular disease risk and 756
 Sudden adult death syndrome (SADS) 882
 Sudden unexpected death in infancy (SUDI) 882, 882*t*
 Sugars, dietary 181, 220–221
 cardiovascular disease risk and 756
 dental caries and 208
 digestion and absorption 221–222, 221*f*
 Suicidal poisoning 788
 Sulfonylureas 314–315
 adverse effects 314–315
 hypoglycaemia 340–341, 346
 indications and clinical usage 315
 mechanism of action 314
 Sulfuric acid, daily production and elimination 68*t*
 Sunlight avoidance, porphyria management
 congenital erythropoietic porphyria 545
 erythropoietic protoporphyria 547
 Superovulation regimens 441
 Supplements, dietary 208–209
 Surfactant 485
 fetal synthesis 444–445, 485
 immature in pre-term infants 485
 neonatal administration 486
 Surgery
 bariatric. *See* Bariatric surgery
 perioperative management of diabetic patients 331
 pituitary gland. *See* Pituitary surgery
 thyroid gland. *See* Thyroid surgery
 Surviving Sepsis Campaign (SSC) 410
 SvO₂ (mixed venous oxygen saturation) 90
 measurement 91
 Sweat
 fluid composition 34*t*
 phosphate content 110

- Sweating
calcium loss 95
excessive, potassium depletion 53
magnesium loss 118
- Sweat test, cystic fibrosis 863
- Swyer syndrome 425
- Symmetrical polyneuropathies 691
- Sympathetic-adrenomedullary system,
activation in acute hypoglycaemia 320
failure 320
- Sympathomimetic syndrome 790*t*
- Synacthen test 354, 355, 358, 372
- Syndrome of inappropriate antidiuretic
hormone secretion (SIADH) 48–49,
49*b*
- Syndrome X. *See* Metabolic syndrome
- Synovial fluid 636–637
examination 644
potassium measurement, post-mortem
879–880
- Synovial joint 636–637, 637*f*
- Synthetic analogues, inherited metabolic
disease management 478
- Syrup of ipecacuanha 794
- Systematic reviews 25
- Systemic inflammatory response syndrome
(SIRS) 165, 409, 409*b*
- Systemic lupus erythematosus (SLE) 587,
639*t*
investigation 587, 587*t*, 644
- T**
- Tachypnoea, transient of the newborn 486
- Tacrolimus 600*t*
therapeutic drug monitoring 784
- Tadalafil 459
- Tamm–Horsfall glycoprotein 131, 154–155
- Tamoxifen
as cause of hypercalcaemia 103*t*
hyperlipidaemia association 732
- Tandem mass spectrometry (TMS)
acylcarnitine profiling 473
amino acid analysis 472
- Tangier disease 696, 722*t*, 728–729
- Tanner staging 415, 415*f*, 416
- Taq polymerase 854–855, 854*f*
- Tartrate-resistant acid phosphatase 613
- 'TATA' box 846
- Tau protein 665, 678
- Tau-transferrin 666
- Tay–Sachs disease 700
- T cell receptor (TCR) 561, 566, 566*t*, 569
immunoglobulin receptor *vs.* 566*t*, 569
- Tea 757
- Technical factors, affecting test results 7
- Teicoplanin, therapeutic drug monitoring
781
- Telopeptides, collagen 612–613, 612*f*, 630,
632
- Tendon xanthomata, in familial
hypercholesterolaemia 726, 726*f*, 727*f*
- Teratogens, retinol 185
- Teratoma
seminoma *vs.* 832
testicular 453. *See also* Germ cell tumours
- Teriparatide 618
- Testes 451–453
autonomously hyperfunctioning 429
development 412–413, 413*f*
failure to descend (cryptorchidism) 455
function. *See* Testicular function
germ cell tumours. *See* Seminoma
Leydig cells. *See* Leydig cell(s)
Sertoli cells. *See* Sertoli cell(s)
size, pubertal changes 415*f*, 416
- Testicular cancer 453
germ cell tumours. *See* Seminoma
tumour markers 841–842
- Testicular function 451
in chronic kidney disease 146
defective hormone synthesis and hormone
receptor defects 456
effects of endocrine disrupting chemicals 453
evaluation 454–455
hypothalamo–pituitary control 452–453,
453*f*. *See also* Spermatogenesis;
Testosterone
- Testosterone 262
biological actions 451–452, 452*t*
concentrations, plasma
age-related changes 457
borderline, interpretation 368
in cirrhosis 262, 262*t*
elevated in females, hirsutism and 441
pituitary function assessment 352–353
deficiency
age-related 457
effect on calcium/bone metabolism 99
in females 439
hirsutism and 441
fetal 412–413, 451–452, 452*t*
secretion 452
in puberty 414
synthesis 452
fetal testes 412–413
human chorionic gonadotrophin
stimulation test 422, 423*f*
in XY disorders of sex development. *See*
XY disorders of sex development
- Testosterone replacement therapy 371,
456–457
- Tests
comparison of results. *See* Comparison of
results
factors affecting results 7–13
analytical 10–12
postanalytical 12–13
preanalytical 7–10
interpretation 13–16
normal *vs.* abnormal results 13
predictive value. *See* Predictive value of
tests
requests 6, 7*t*
specific uses of 2–5. *See also specific disorders.*
specific tests
- Tetracosactide test 354, 355, 358, 372
- Tetrahydrobiopterin deficiency 698
- Tetrahydrocannabinol 802
- Tetrahydrofolate (THF) 521, 522*f*
- Thalassaemia(s) 550, 551, 552–554
 α thalassaemia 552–553, 552*t*, 558
 β thalassaemia 552, 553–554, 554*b*, 554*t*,
557–558
epidemiology 550, 551*f*, 552
laboratory diagnosis 557–559, 557*t*, 558*f*
- Thallium, antidote 794*t*
- Thanatochemistry 875*b*. *See also*
Post-mortem biochemistry
- Thelarche 415
premature 429
- Theophylline
poisoning 792*t*, 803
management 803
therapeutic drug monitoring 783
- Therapeutic biomarkers 776
- Therapeutic diets 208–209, 209*t*. *See also*
Dietary management
- Therapeutic drug monitoring (TDM) 4,
767–786
aims 767
analytical methods 774, 774*t*
clinical interpretation 774–775
clinically useful, criteria 770*b*
continuing education 776
drug interactions 771, 772*t*
drugs measured 769–771, 770*b*
effective, criteria 771, 771*b*
accurate analysis 774, 774*t*
accurate patient information 772, 772*t*
appropriate clinical question 771–772
appropriate sample 772–774
effective action taken 774–775, 775*f*
relevant clinical interpretation 774–775
- individual drugs 777–785
analgesic/anti-inflammatory drugs 777
antiarrhythmics and cardiac glycosides
777–778
anticonvulsants 778–780
antidepressants 780
antifungal drugs 781–782
antimicrobials 781–782
antineoplastic drugs 782–783
antipsychotic drugs 780
antiretroviral drugs 782
antitubercular drugs 782
bronchodilator drugs 783
immunosuppressants 783–785
opiate and opioid drugs 785
- integrating information 777
patient information requirements 772,
772*t*
pharmacodynamic monitoring 776–777
pharmacogenetic studies 776
point-of-care testing 775
quality assurance 776
reporting 775
saliva 773
sampling 772–774
service provision 775–776
staff 775
therapeutic biomarkers 776
toxicity 771
turnaround time 775
units 775–776
use of 771–775
- Therapeutic index 770
- Thevetia 798, 806
- Thiamin (vitamin B₁) 187, 694
laboratory-based assessment 197
- Thiamin (vitamin B₁) deficiency 187, 197,
212, 677
manifestations 187
encephalopathy 187, 686
peripheral neuropathy 187, 694
- Thiazide diuretics
abuse 35
as cause of hypercalcaemia 103*t*
congenital nephrogenic diabetes insipidus
management 43
effect on biochemical variables 8*t*
hypertension management in diabetes
mellitus 309
low osmotic load hyponatraemia and 50
- Thiazide-like diuretics, hypercalcaemia
treatment 176
- Thiazolidinediones 315–316
adverse effects 316
hypoglycaemia risk 341
mechanism of action 315–316
- Thin layer chromatography (TLC), inherited
metabolic disease investigation
470–471, 472
- Thioctic acid 692
- Thiopurine S-methyltransferase (TPMT) 862
- Thirst
in acute hypernatraemia 38
in chronic kidney disease 147

- Thirst (*Continued*)
 compulsive water drinking 41, 681
 in diabetes insipidus 39–40, 40f
 excessive. *See* Polydipsia
 non-osmotic control 32
 osmoregulation 31–32, 31f
 pathological 41
 water deficiency with 44
 water deficiency without 44–46, 45f
- Thrombin 507, 508f
- Thrombin time (TT) 508, 509t
- Thrombocytopenic purpura (TTP) 529
- Thrombogenesis, cardiovascular disease risk and 753
- Thrombomodulin 408–409
- Thrombosis
 risk in nephrotic syndrome 159. *See also* Venous thromboembolic disease
- Thrombospondins 607, 607t
- Thromboxane(s) 705
 stress response 406
- Thromboxane A₂ 705, 706f
- Thymus 562
- Thyocytes 374, 374f
 biochemical processes 375–376, 375f
- Thyroglobulin (Tg) 375, 375f
 antibodies to 386–387, 842
 measurement 386
 analytical and reporting requirements 826t, 842
 post-mortem biochemistry 879t
 as tumour marker 823t
 thyroid cancer monitoring 400, 824t, 842
- Thyroidal antigens, autoantibodies to 386–387, 583, 601t
- Thyroid cancer 399–400, 399b, 842
 diagnosis 391f, 399–400, 400b, 842
 medullary. *See* Medullary thyroid carcinoma
 monitoring 842
 screening 842
 treatment 400
 tumour markers 400, 812, 842
 analytical and reporting requirements 842
- Thyroid crisis 388
- Thyroid disease 373–374
 autoimmune 386, 583, 601t
 hypertension and 759
 hyperthyroidism. *See* Hyperthyroidism
 hypothyroidism. *See* Hypothyroidism
 inflammation. *See* Thyroiditis
 investigations. *See* Thyroid function tests
 neoplasia. *See* Thyroid cancer
 postpartum 399
 screening 401
 secondary 385
 subclinical 385. *See also specific diseases*
- Thyroidectomy
 Graves disease treatment 392
 thyroid cancer treatment 400
 medullary thyroid cancer 813
- Thyroid function 374
 abnormalities
 in chronic kidney disease 146–147, 146b.
See also Thyroid disease
 assessment 382–388
 clinical evaluation of thyroid status 382–383, 382t
 following pituitary surgery 359
 imaging techniques 387–388
in vitro tests. *See* Thyroid function tests
 depression and 678, 680
 extrathyroidal factors affecting 379–382
 age 379
 drugs 381–382, 381t
 non-thyroidal illness 380–381, 381f
 pregnancy 379, 385, 445
- Thyroid function tests 383
 free hormone measurements 383
 assay nomenclature 384
 methods 383
 theoretical considerations 383
 validity of commercial methods 383–384
 interpretation 382t, 385
 miscellaneous tests 386
 in psychiatric disorders 680
 anxiety disorders 677
 depression 678, 680
 post-traumatic stress disorder 679
 reference ranges 385–386
 selective use 384
 significant changes 385–386
 total hormone measurements 384
 TSH measurement. *See* Thyroid stimulating hormone (TSH), measurement
- Thyroid gland
 disease. *See* Thyroid disease
 embryology 374
 follicular cells. *See* Thyrocytes
 function. *See* Thyroid function
 imaging 387–388
 in multiple endocrine neoplasia type 2 813. *See also* Thyroid scintiscanning
 nodules 399–400
 palpation 382–383
 physiology 374
 surgery. *See* Thyroid surgery
- Thyroid hormone(s) 373–374
 in anorexia nervosa 207
 biological actions 374–375, 375b
 catabolism 377
 in chronic kidney disease 146b
 classic feedback regulation 378, 378f
 deiodination 377
 effect on calcium/bone metabolism 99, 374
 effects on metabolic indices 375, 375b
 entry into tissues 377
 free, measurement. *see under* Thyroid function tests
 free hormone hypothesis 376–377
 maternal–fetal transfer 380
 nuclear action 377
 receptors 377
 reference ranges 385–386
 resistance 377, 385, 400–401
 thyroid stimulating hormone-secreting adenomas *vs.* 400–401, 401t
 secretion 375–376, 375f
 control 378–379
 effect of radiotherapy 819
 significant changes 385–386
 storage 375–376, 375f
 stress response 405
 structure 374f
 synthesis 374, 375–376, 375f
 control 378–379, 378f, 379f
 iodine and 376
 total, measurement 384
 transporters 376, 377
 transport in blood 376. *See also* Thyroxine (T₄); Tri-iodothyronine (T₃)
- Thyroid hormone replacement therapy 371, 397–398
 drugs affecting 381–382, 398
 in pregnancy 380
- Thyroiditis 398–399
 De Quervain 399
 Hashimoto. *See* Hashimoto disease
 hyperthyroidism-producing 398–399
 postpartum 399
 Riedel 399
- Thyroid peroxidase (TPO) 375, 375f, 376
 antibodies 386
 measurement 583
 transplacental passage 380
 inhibition by drugs 376
- Thyroid scintiscanning 387–388, 391f
 Graves disease 390, 391f
 thyroid cancer 391f, 399–400
 medullary 813
 toxic adenoma 391f, 392
 toxic multinodular goitre 391f, 392
- Thyroid stimulating hormone (TSH) 376, 378
 age-related changes 379
 deficiency 352, 360, 369
 in hyperthyroidism 382t, 389
 subclinical 394
 measurement 383
 interpretation of results 382t, 385
 misleading results 385
 reference ranges 385
 significant changes 385–386
 thyroid cancer monitoring 842
 pregnancy-related changes 380, 385
 receptor antibodies 387, 583
 release, regulatory mechanisms 351, 379, 379f
 classic feedback regulation 378, 378f
 stress response 405
 suppression, thyroid cancer management 400, 400b
- Thyroid stimulating hormone-secreting adenomas (TSHomas) 366–367, 385, 390t, 392–393
 diagnosis 392–393
 thyroid hormone resistance *vs.* 400–401, 401t
 treatment 393
- Thyroid surgery
 Graves disease treatment 391, 392
 hypothyroidism following 396
- Thyroid tissue, ectopic 390t, 393
- Thyrotoxicosis
 diabetes association 294, 295–296
 eye signs 390t
 hypokalaemic periodic paralysis with 53
- Thyrotoxicosis factitia 390t, 393
- Thyrotrophin releasing hormone (TRH) 351, 378, 378f
 stimulation of prolactin secretion 440
 tests 356, 386
 transplacental passage 380
- Thyroxine (T₄) 351
 age-related changes 379
 extrathyroidal deiodination 377
 free, measurement 383
 methods 383
 reference ranges 385
 significant changes 385–386
 theoretical considerations 383
 interpretation of thyroid function tests 382t
 pregnancy-related changes 380
 reference ranges 385
 replacement therapy. *See* Thyroid hormone replacement therapy
 secretion 375f, 376
 stress response 405
 structure 374f
 synthesis 374, 375, 375f
 total, measurement 384
 reference ranges 385
 significant changes 385–386
 transporters 377
 transport in blood 376. *See also* Thyroid hormone(s)

- Thyroxine-binding globulin (TBG) 376
 abnormal concentrations, causes 384, 384b
- Thyroxine-binding prealbumin (TBPA) 376
- Tibolone 450
- Tics 699
- Time-dependent changes, effect on test results 8–9
- Tissue factor (TF) 507
- Tissue inhibitor of metalloproteinase 1 (TIMP-1), serum tests for hepatic fibrosis 244
- Tissue oxygenation 65, 87–92
 effects of acidosis 76
 factors affecting 90t
 haemoglobin role in oxygen transport 88–89, 88f
 impaired. *See* Hypoxia, tissue
 oxygen delivery 89, 91, 91f
 measurement 90–91
 oxygen uptake
 into blood. *See* Oxygen uptake
 into tissues 90
 pulmonary function 87–88
- Tissue polypeptide antigen (TPA) 823t
- Tissue protective receptors (TPR) 407
- Tissue sampling techniques, fetal 444, 444f, 444t
- T lymphocytes 563t, 566
 antigen receptor. *See* T cell receptor (TCR)
 clonality 561
 deficiency
 associated infections 576, 577t
 primary immunodeficiencies 578
- TmP/GFR (tubular maximum for phosphate/glomerular filtration rate) 111
 estimation 111, 114, 123, 123f
 factors regulating 111–112, 112t
 in hyperphosphataemia 113–114
 in hypophosphataemia 116
- Tobramycin, therapeutic drug monitoring 781
- α -Tocopherol 185, 185f. *See also* Vitamin E
- Toluene abuse 803
- Tolvaptan 51
- Tonicity 27
- Total allowable error 12
- Total carbon dioxide (TCO₂), measurement 73
- Total iron binding capacity (TIBC) 520
- Toxic adenoma 392
 diagnosis 391f, 392
 treatment 392
- Toxic metabolites, in inherited metabolic diseases
 blockage of site of action 482
 reduction strategies 481–482
 removal strategies 482–483
- Toxic multinodular goitre 392
 clinical features 392
 diagnosis 391f, 392
 pathogenesis 390t
 treatment 392
- Toxicodynamics 787
- Toxicology screen 792t
- Toxins
 interaction with biological systems. *See* Poisoning
 uraemic 143, 143b, 145–146
- TP53 868
- Trabecular bone 604
- Trace elements 184, 190–192
 laboratory-based assessment 198–199
 parenteral nutrition 211. *See also specific trace elements*
- Trait, definition 872
- Transaminases. *See* Aminotransferases
- Transcellular fluid 27
- Transcobalamins 521
- Transcription 846, 847f
 definition 872
- Transcription errors 12
- Transcription factors, in erythropoiesis 516
- Transcriptomics 871
- Transfection, definition 872
- Transferrin 242, 519
 carbohydrate-deficient 242–243, 258
 cerebrospinal fluid 662–663, 663f, 663t
 in haemochromatosis/iron overload 242, 265–266
 liver function tests 242–243
 plasma concentration, in malnutrition 194–195
 saturation 520
 serum concentration, measurement 520
 serum receptor, measurement 520
- Transformation, definition 873
- Transfusion, blood. *See* Blood transfusion
- Transfusion reaction, investigation 512
- Transgenic animal, definition 873
- Transglutaminase 220
- Transient hypogammaglobulinaemia of infancy 579
- Transient paraproteinaemia 596
- Transient tachypnoea of the newborn (TTN) 486
- Transitions 846–847, 848f
- Translation 846, 847f
 definition 873
- Transplantation
 immunological issues 599–600, 599b. *See also specific types*
- Transthyretin (prealbumin) 376
 in assessment of nutritional status 195
 cerebrospinal fluid 663f, 663t, 666
 gene mutation in familial amyloid polyneuropathy 696
 in pregnancy 446
- Transtubular potassium gradient (TTKG)
 calculation 56–57, 62–63
 hyperkalaemia 60–61
 hypokalaemia 56–57
- Transurethral prostatectomy syndrome (TURS) 46
- Transversion 846–847, 848f
- Triazodone overdose 804
- Tremor 698
- Triacylglycerols. *See* Triglyceride(s)
- Tricarboxylic acid cycle 647–649
 glucose metabolism and 278
- Triceps skinfold (TSF) measurement 194
- Tricyclic antidepressants
 poisoning 803–804
 clinical features 803
 management 803–804
 mechanisms 803, 804f
 therapeutic drug monitoring 780
- Trientine hydrochloride, Wilson disease
 treatment 267
- Triglyceride(s) 704
 absorption 224, 711
 cardiovascular disease risk and 752–753
 dietary 182
 digestion 224, 711
 function 703t
 measurement, investigation of lipid disorders 732
 metabolism 234–235, 704
 complete 69
 incomplete 69
- lipolysis in adipose tissue 713. *See also* Lipid metabolism
 plasma concentration. *See* Triglycerides, plasma concentration
 structure 704f
- Triglyceride-rich lipoproteins (TGRLs) 715
- Triglycerides, plasma concentration
 analytical goals 11t
 analytical variation 11t
 biological variation 11t, 732
 critical difference 15t
 effect of food intake 9
 in nephrotic syndrome 160, 160t
 in obese patients 8
- Tri-iodothyronine (T3) 351
 age-related changes 379
 extrathyroidal production 377
 free, measurement 383
 methods 383
 reference ranges 385
 significant changes 385–386
 theoretical considerations 383
 pregnancy-related changes 380
 reference ranges 385
 secretion 375f, 376
 stress response 405
 structure 374f
 synthesis 374, 375, 375f, 377
 total, measurement 384
 reference ranges 385
 significant changes 385–386
 transporters 377
 transport in blood 376. *See also* Thyroid hormone(s)
- Trimipramine, therapeutic drug monitoring 780
- ^{13/14}C-Triolein breath test 225
- Triple acid–base disorders 87
- Triple phosphate stones 177
- Triple test 443, 444
- Trisomies 848–849, 852
- Trisomy 21. *See* Down syndrome
- Troglitazone 315
- Trophoblastic tumours. *See* Gestational trophoblastic neoplasia (GTN)
- Tropical diabetes 293–295
- Tropomyosin 740, 740f
- Troponin(s) 740, 740f
 as biomarkers of acute myocardial damage 745–746, 745f
 high-sensitivity troponins 746
 elevated, causes other than acute coronary syndrome/heart failure 746–747, 747b
 post-mortem biochemistry 879t, 880
- Troponin C 740, 740f
- Troponin I 740, 740f
 as biomarker of acute myocardial damage 745–746, 745f
- Troponin T 740, 740f
 as biomarker of acute myocardial damage 745–746, 745f
- True hermaphroditism 416, 417t, 423–425.
See also Disorder of sex development (DSD)
- True negatives 16, 16t, 17t
- True positives 16, 16t, 17t
- Trypsin 223
 pancreatic function testing 217–218
- Trypsinogen 217–218
- Trypsate 582
 post-mortem biochemistry 879t, 880
- Tryptophan 809f
 in carcinoid syndrome 808–809
 catabolism, nicotinamide formation 188
- Tuberculosis, therapeutic drug monitoring 782

- Tuberous xanthomas, in remnant hyperlipoproteinaemia 726, 726f
- Tubular necrosis, acute. *See* Acute tubular necrosis (ATN)
- Tubular proteinuria 160–162
glomerular proteinuria *vs.* 160, 160t, 161
renal disorders associated with 160–161
- Tubuloglomerular feedback 128–129
- Tubulointerstitial nephritis, causes 160–161, 161b
- Tumoral calcinosis 113–114
- Tumour(s)
endocrine sequelae 818–820, 819t
loss of heterozygosity 868, 868f
markers. *See* Tumour markers; Cancer; Malignancy. *specific types*
- Tumour-associated glycoprotein antigen (TA90-IC) 837
- Tumorigenesis, multiple endocrine neoplasia type 1 811
- Tumour lysis syndrome 114, 115, 816
- Tumour markers 821–843
analytical requirements 826
characteristics 822t, 823t
choice of test 825
definition 821
evaluation of clinical utility of 821–825
International Reference Preparations (IRP) 826t
International Reference Reagents (IRR) 826t
International Standards (IS) 826t
key points for use 843b
management of specific cancers 827–842
bladder cancer 827
breast cancer 827–829
cancers of unknown primary origin 842
cervical cancer 829
colorectal cancer 829–831
gastric cancer 831
gastrointestinal stromal tumours 831–832
germ cell tumours 832–834
gestational trophoblastic neoplasia 834–835
hepatocellular carcinoma 264, 835–836
lung cancer 836–837, 836t
melanoma 837
neonatal and paediatric tumours 837–838
ovarian cancer 838–840
pancreatic cancer 840
prostate cancer 840–841
testicular cancer 841–842
thyroid cancer. *See* Thyroid cancer
misleading results, causes 826, 826b
pre-analytical requirements 825
reasons for requesting 824t, 825
reporting of results 827, 833f
- Tumour necrosis factor superfamily 573t
- Tumour necrosis factor α (TNF α) 406
- Tumour suppressor genes 867–869
definition 873
- Tunica adventitia 741, 741f
- Tunica intima 741, 741f
- Tunica media 741, 741f
- Turner syndrome 431
management 431
- Two-cell theory 438
- Tyrosinaemia 269, 269f
- Tyrosinaemia type I 269, 462–463, 481–482, 495–496
management 481–482, 482f, 495
neonatal 495
- Tyrosinaemia type II 269
- Tyrosinaemia type III 481–482, 482f
- Tyrosine, metabolism 269f
- U**
- Ubiquinol 656
- Ubiquinone, in measurement of respiratory chain complex activity 656
- UDP-glucuronosyltransferase 862
bilirubin metabolism 493
in Crigler–Najjar syndrome 246, 494
in Gilbert syndrome 246
- Ulcerative colitis
faecal calprotectin concentrations 225–226. *See also* Inflammatory bowel disease (IBD)
- Ultracentrifugation, lipoprotein separation 733–734
- Ultrafiltration methods, free thyroid hormone measurement 383
- Ultrasonography
evaluation of internal anatomy in disorder of sex development 418
polycystic ovary syndrome 435f
screening for fetal malformation 443
thyroid gland 387
in multiple endocrine neoplasia type 2 813
- Undernutrition 200
therapeutic diets 209t. *See also* Malnutrition
- Universal donors 510
- Universal recipients 510
- Unstable angina 738, 739f
- Upstream, definition 873
- Uraemia, sodium deficiency 35
- Uraemic acidosis 78–79
- Uraemic encephalopathy 686
- Uraemic hypoglycaemia 344–345
- Uraemic syndrome 129, 142–144
clinical features 142b, 143
definition 141
- Uraemic toxins 143, 143b, 145–146
- Urate
age-related changes 8
analytical variation 11t
biological variation 11t
critical difference 15t
in hereditary renal hypouricaemia 173–174
in hyponatraemia 50
renal tubular handling 169t
- Urate oxidase (uricase) 642
recombinant 642, 816
- Urea
disposal 235, 235f
plasma concentration. *See* Urea, plasma concentration
renal tubular handling 169t
role of 129
synthesis 69–70
hydrogen ion production 68t
vitreous humour, post-mortem biochemistry 879t
- Urea, plasma concentration
analytical goals 11t
analytical variation 11t
biological variation 11t
in chronic kidney disease 143
critical difference 15t
effect of food intake 9
hyponatraemia 50
in pregnancy 445
renal function assessment 135
- Urea cycle 235, 463f
- Urea cycle defects 463f
investigation
plasma ammonia measurement 469, 470
urinary orotic acid analysis 473
management
drugs 482–483, 483f
- liver transplantation 480
metabolic load reduction 481
presentation at puberty 464–465
presentation in adulthood 465
presentation in neonates 462, 470
- Urea kinetic modelling (UKM) 149
- Ureterosigmoidostomy 54, 55
- Ureterostomy, cutaneous 54
- Uric acid 639–640
stones 175t, 177
synthesis 640f
- Urinalysis
renal function assessment 130–131. *See also* Urine
- Urinary acidification test 179
- Urinary frequency 129
- Urinary reducing substances, in inherited metabolic diseases 470–471
- Urinary steroid profiling, in disorder of sex development 420
- Urinary tract obstruction 34
- Urine
abnormal colouration 130, 130b
acidification 70–71, 72f
albumin. *See* Albumin
amino acids, analysis 472, 472b
appearance 130–131
casts 131
dip-sticks 165
diversion from bladder 54
excessive output. *See* Polyuria
formation 31
glucose
measurement 299. *See also* Glycosuria
glycosaminoglycans, analysis 474
ketone testing. *See* Ketone testing
normal colouration 130
organic acid analysis 472–473, 473b
orotic acid analysis 473
pH 70–71, 131
protein
measurement 165–166, 166f. *See also* Protein, urinary excretion
sediment 131
specific gravity 131
turbidity 130–131
- Urine collection
investigation of stone formers 178
nocturnal polyuria 43–44
proteinuria 165
- Urine osmolality 31, 129, 131
chronic kidney disease 147
hyponatraemia 35, 50
neonates 487
sodium deficiency 35
- Urobilinogen
enterohepatic circulation 237–238, 238f
in pre-hepatic jaundice 245
urinary 245
- Urolithiasis 174. *See also* Renal calculi
- Uromodulin (Tamm–Horsfall glycoprotein) 131, 154–155
- Uromodulin storage disease 154–155
- Uroporphyrinogen decarboxylase (UROD) 533–534, 535f
activity in hepatoerythropoietic porphyria 545
deficiency 544. *See also* Porphyria cutanea tarda (PCT)
- Uroporphyrinogen III 533–534, 535f
- Uroporphyrinogen synthase (UROS), deficiency. *See* Congenital erythropoietic porphyria (CEP)
- Ursodeoxycholate 236
- User satisfaction surveys, laboratory services 23, 23b
- Uterine changes, menstrual cycle 436

- Uterus, effects of radiotherapy 820
 UV therapy, erythropoietic protoporphyria 547
- V**
- Vacuolar protein sorting 10 protein (VPS10P) 721
 Valine, metabolism 480f
 Valproate
 hepatotoxicity 682
 therapeutic drug monitoring 779
 Vancomycin, therapeutic drug monitoring 781
 Vaptans 51, 814
 Vardenafil 459
 Variable number of tandem repeats (VNTRs)
 859, 859f
 definition 873
 Variegate porphyria (VP) 534t, 537–538, 537t
 homozygous 542
 molecular genetics 534–537, 537t
 pathophysiology 538
 screening 541t. *See also* Porphyria(s), acute
 Vasa recta 128
 Vascular endothelial growth factor (VEGF)
 antagonists, macular oedema
 treatment 325–326
 Vaso-occlusion, in sickle cell anaemia 555
 Vasopressin. *See* Arginine vasopressin (AVP)
 Vasopressinase 41
 Vasopressinase-related diabetes insipidus 41
 Vasopressin test 63
 Vectors
 definition 873
 gene therapy 870
 Vegetable consumption, cardiovascular
 disease risk and 757
 Vegetarians, cardiovascular disease risk 757
 Venesection
 haemochromatosis treatment 266, 643
 porphyria cutanea tarda treatment 544
 Venlafaxine overdose 804
 Venous thromboembolic disease
 combined oral contraceptive-associated
 risk 448
 D-dimer measurement 508
 hormone replacement therapy-associated
 risk 449
 Ventilation, alveolar 87
 Ventilation–perfusion imbalance 87–88, 89
 Ventilatory support. *See* Respiratory support
 Ventricles, cerebrospinal fluid 661
 Versican 606–607
 Vertical sleeve gastrectomy (VSG) 206, 206f
 Very long chain fatty acids (VLCFAs) 704
 in inherited metabolic disease 462, 465
 analysis 474
 Very long chain hydroxyl-acyl-CoA
 dehydrogenase deficiency (VLCADD)
 465–466, 473
 Very low density lipoprotein (VLDL) 704,
 706–707, 708
 characteristics 708t
 metabolism 711–712, 713–714
 in nephrotic syndrome 160, 160t
 in type 2 diabetes 290
 Vesicant poisoning 790t
 Villous adenoma of rectum 53–54
 VIPoma 227
 Viral hepatitis, acute. *See* Hepatitis, acute
 Virilism 441–442
 Virilization, external genitalia
 differential 418f
 females 413–414, 419, 423–425
 Vision, role of vitamin A 184–185
 Visual field assessment, patients with
 pituitary disease 360
- Vitamin(s) 184–192
 definition 96
 fat-soluble 184–187, 192, 224
 laboratory-based assessment 196–198
 water-soluble 187–190, 192. *See also specific vitamins*
- Vitamin A 184–185, 717
 deficiency 185, 197
 function 184–185, 703t
 laboratory-based assessment 197
 structure 184f
 toxicity 103t, 185
- Vitamin B₁. *See* Thiamin (vitamin B₁)
 Vitamin B₂. *See* Riboflavin (vitamin B₂)
 Vitamin B₃. *See* Niacin
 Vitamin B₅. *See* Pantothenic acid
 Vitamin B₆ 188, 188f
 deficiency 188, 198, 694
 hyperoxaluria and 176
 laboratory-based assessment 198
 toxicity 694
- Vitamin B₁₂ 521
 absorption 521
 metabolism 476f, 521
 disorders 476–477
 requirements 519t, 521
 serum concentration, measurement
 523–524
 sources 519t, 521
 status, laboratory determination 523–524
 structure 521, 523f
- Vitamin B₁₂ deficiency 217, 219, 501, 521
 causes 523
 features 686
 encephalopathy 686
 peripheral neuropathy 693
 spinal cord degeneration 688–689
 laboratory investigation 523–524
 in malignant disease 816t
 serum folate concentration 521
 treatment 693. *See also* Pernicious anaemia
- Vitamin C 189
 deficiency 189, 198
 dietary sources 189
 excess intake 189
 function 189
 laboratory-based assessment 198
- Vitamin D 96–98
 actions 97–98
 circulating metabolites 97, 97t
 function 703t
 laboratory-based assessment 197, 621
 measurement of metabolites 97t, 98
 receptor 97–98
 synthesis and metabolism 96–97, 97f
 synthetic analogues 98
 hyperparathyroidism treatment 627
 toxicity 103, 197, 627
- Vitamin D₂ 96
 Vitamin D₃ 96
 Vitamin D deficiency 197, 615, 617
 haemolysis in neonates 529
 in nephrotic syndrome 159–160
 osteomalacia 620–621
 in primary hyperparathyroidism 628–629
 treatment 109
 responses 622. *See also* Vitamin D
 supplementation
- Vitamin D-dependent rickets (VDDR) 621
 treatment 622
- Vitamin D supplementation 98
 in hyperparathyroidism 628–629
 in nephrotic syndrome 159–160
 in osteopenia of prematurity 492
 in osteoporosis 617
 in phosphopenic osteomalacia 623
- Vitamin E 185–186
 atherogenesis and 743
 deficiency 185, 197, 689, 694
 ataxia with 185, 699–700
 function 703t
 laboratory-based assessment 197
 structure 185f
 supplementation 185
- Vitamin K 186–187
 antagonists 186f
 deficiency 186–187, 197
 dietary 186
 function 186, 703t
 laboratory-based assessment 197
 neonatal administration 186–187
 structure 186f
 toxicity 187
- Vitreous humour, post-mortem biochemistry
 879–880, 879t
- Vitronectin 607, 607t
 Volatile substance abuse 792t, 803
 von Gierke disease 277
 von Hippel–Lindau syndrome 813
- W**
- Waardenburg syndrome 851
 Waist circumference 194
 Waldenström macroglobulinaemia 591t, 595,
 595t, 693
 'Washout' phenomenon 91
 Wasting syndrome, cancer patients 818
 Water 27
 balance
 in chronic kidney disease 147
 in hyperthyroidism 389
 in hypothyroidism 395
 body content 27, 28t
 pre-term *vs.* term infants 487, 487f
 compulsive drinking 41, 681
 deficiency
 with thirst 44
 without thirst 44–46, 45f
 deficit calculation in hypernatraemia 62
 fractional excretion of 878, 878b
 intake control 31–32
 intoxication, estimation of sodium required
 62
 intracellular fluid and 30–32
 metabolism, disorders 39–51
 hypernatraemia. *See* Hypernatraemia
 hyponatraemia. *See* Hyponatraemia
 polyuria. *See* Polyuria
 renal output, control 30–31
 restriction, hyponatraemia management
 51
 retention
 ascites 259–260
 in nephrotic syndrome 159
 Water channels (aquaporins) 27, 30, 129
 Water deprivation test 42, 63, 357, 372
 interpretation 42, 63, 372
 protocol 42, 63, 63t, 372
 Water drinking
 compulsive 41, 681. *See also* Thirst
 Water load test 64, 64t
 interpretation 64
 Water-soluble vitamins 187–190, 192. *See also specific vitamins*
- Watery diarrhoea–hypokalaemia achlorhydria (WDHA) syndrome 53–54
 Weaning, inherited metabolic disease
 presentation 464
 Wegener granulomatosis 639t
 Weight, assessment of nutritional status
 193–194, 193t

- Weight gain
 growth hormone deficiency 203. *See also*
 Obesity
- Weight loss 193, 205
 amenorrhoea and 369, 440
 in anorexia nervosa 207
 in cancer patients. *See* Cancer cachexia
 dietary approaches 205
 in hyperthyroidism 388
 pharmacological approaches 205
 physical exercise 205
 surgical approaches. *See* Bariatric surgery
- Wermer syndrome. *See* Multiple endocrine
 neoplasia (MEN), type 1
- Wernicke–Korsakoff syndrome 187, 686
- Western blot, definition 873
- Whipple's triad 227, 334
- White blood cells (WBCs) 515, 562–566, 564*f*
 in cerebrospinal fluid 661, 662
 count 498, 498*t*, 562, 563*t*, 602
 differential 498–499, 499*f*, 563*t*
 low 504
 raised 503–504
 function 503
 morphology 503–504, 504*f*, 564*f*
 abnormal 502*f*, 504, 505*f*
 in redistribution hyperkalaemia 58
 in redistribution hypokalaemia 52
 stress and 407
 urinary casts 131. *See also specific types*
- Wild type, definition 873
- Williams syndrome 491
- Wilson disease 191, 266–268, 496, 698–699
 diagnosis 243, 254*t*, 255, 266–267, 267*t*
 long-term management 267
- Wolff–Chaikoff effect 376
- Wolffian ducts 413, 451
- Wolfram syndrome 39–40
- Wolman disease 717
- World Health Organization (WHO),
 diagnostic criteria for diabetes
 mellitus 282, 283*b*, 283*t*, 284*b*
- X**
- Xanthine oxidase 639–640, 640*f*
 deficiency 177–178
- Xanthine stones 177–178
- Xanthinuria 177–178
- Xanthochromia 661, 667
- Xanthomata 723*t*
 in acquired hyperlipidaemia 729
 eruptive, in chylomicronaemia syndrome
 725, 725*f*
 in remnant hyperlipoproteinaemia 726, 726*f*
 tendon, in familial hypercholesterolaemia
 726, 726*f*, 727*f*
- X chromosome, gene mapping 845, 845*f*
- Xeroderma pigmentosum 868–869
- X-linked agammaglobulinaemia 578
- X-linked dominant protoporphyria (XLDPP)
 533, 534*t*, 546–547
 biochemical findings 536*f*, 537*t*
 chronic complications and their
 management 547
 molecular genetics 537*t*
 symptoms and signs 546–547
 treatment 547
- X-linked hyper-IgM syndrome 578
- X-linked hypophosphataemic rickets 112,
 622, 623
- X-linked inheritance 467, 468*f*, 850, 850*f*
- X-linked nephrolithiasis 172
- XX disorders of sex development 423, 424*t*
 androgen excess 423–425, 424*t*
 46XX ovotesticular DSD 423–425
 46XX testicular DSD 423–425. *See also*
 Disorder of sex development (DSD)
- XX males 412–413
- XY disorders of sex development 419–420,
 424*t*, 426*t*, 456
 evaluation of internal anatomy 418
 investigations in adolescents 419–420
 with low testosterone and high steroid
 precursor concentrations 425–427,
 426*t*
 with low testosterone and low precursor
 concentrations 425, 426*t*
 with normal testosterone, normal
 precursor and low DHT
 concentrations 427
 with normal testosterone, normal
 precursor and normal DHT
 concentrations 427–428
 physical examination 417. *See also* Disorder
 of sex development (DSD)
- XY females 412–413
- Xylose absorption test 222
- Y**
- Y chromosome 412–413, 451
- Yolk sac tumours 838
- Z**
- Zellweger syndrome 462, 495
- Zieve syndrome 529
- Zinc 190
 deficiency 190, 198
 dietary sources 190
 functions 190
 laboratory-based assessment 198
 toxicity 190
 Wilson disease treatment 267
- Zoledronate 617, 618, 631*f*, 632
- Zollinger–Ellison syndrome 55, 216, 811